istat
working
papers

N. 3
2013

Il trattamento delle variabili testuali nel 15° Censimento generale della popolazione

A cura di Stefania Macchia e Simona Mastroluca

Istat
Istituto Nazionale
di Statistica

istat working papers

N. 3
2013

Il trattamento delle variabili testuali nel 15° Censimento generale della popolazione

A cura di Stefania Macchia e Simona Mastroluca

Istat

Istat Working Papers

Il trattamento delle variabili testuali nel
15° Censimento generale della popolazione

N. 3/2013

ISBN 978-88-458-1748-9

Istituto nazionale di statistica
Servizio Editoria
Via Cesare Balbo, 16 – Roma

Il trattamento delle variabili testuali nel 15° Censimento generale della popolazione[1]

A cura di Stefania Macchia e Simona Mastroluca

Sommario

Il 9 ottobre 2011 è la data di riferimento del 15° Censimento generale della popolazione e delle abitazioni.
Numerose sono state le innovazioni metodologiche introdotte al fine di semplificare l'impatto organizzativo sulle amministrazioni pubbliche, diminuire il carico statistico sui rispondenti, garantire una tempestiva diffusione delle informazioni rilevate e una qualità sempre più elevata del dato censuario.
Il documento illustra tutte le attività inerenti il trattamento delle variabili testuali, partendo dalla predisposizione delle basi informative per la loro codifica, fino al processo di codifica vero e proprio, progettato in modo tale da massimizzare i casi di individuazione del codice e assicurare la coerenza dei risultati, tenendo in considerazione le due diverse tecniche di rilevazione: compilazione del questionario elettronico e compilazione del modello cartaceo, con successiva acquisizione tramite lettura ottica.
Nel lavoro vengono descritti anche il navigatore delle Professioni e il navigatore delle Attività Economiche, applicazioni utilizzate per la prima volta in occasione di un censimento per aiutare i cittadini a rispondere correttamente a quesiti particolarmente complessi.

Parole chiave: censimento, codifica

Summary

The 15th General Population and Housing Census was held in 2011, with reference date October 9th.
Many methodological innovations have been adopted in order to improve the organizational impact on public administrations, reduce the respondent statistical burden and guarantee a timely dissemination and a higher quality of data.
This document describes all the activities concerning textual variable treatment, from the setting of the informative bases to be used to code, to the coding process itself, designed so as to maximize coding rates and to ensure consistency of results, taking into account the two data capturing techniques used: filling in web questionnaires (CAWI) and paper ones (PAPI) captured by OCR (Optical Character Recognition).
The paper also illustrates the Web navigators for Professions and Economic Activities used for the first time in a Population Census with the purpose of supporting respondents in answering very complex questions.

Keywords: census, coding

[1] Il lavoro è frutto dell'attività congiunta degli autori. In ogni caso, ai soli fini dell'attribuzione, i capitoli 1 e 2 sono da attribuirsi a S. Mastroluca, il cap. 3, le introduzioni ai cap. 4, 5 e 6, i sottoparagrafi 6.2.2, 6.3.2 e il paragrafo 8.1 a S. Macchia, i sottopar. 4.1.1, 4.1.3 e 6.1.1 a T. Clary, i sottopar. 4.1.2 e 4.1.4 a F. Lipizzi e il sottopar. 4.1.5 a Clary/Lipizzi, i sottopar. 4.1.6, 4.1.7 e 6.1.2 a M. Murgia, il sottopar. 4.2.1 a L. Verzicco, i sottopar. 4.2.2, 4.2.3, 4.2.4, 4.2.5 e 6.2.1 a L. Mazza, il par. 4.3 e il sottopar. 6.3.1 a M. Albani, il par. 5.1 a A. Virgillito, i par. 5.2 e 5.3 a L. Tininini, il par. 7.1 a P. Scalisi, il par. 7.2 a F. Gallo, il par. 7.3 a A. Capezzuoli, i par. 8.2, 8.3 e 8.4 a D. Carboni/F. Panicali, il par. 9 a Macchia/Mastroluca. Ha collaborato Giacomo Ricci.

Indice

1. L'esperienza del Censimento 2001

In occasione del 14° Censimento generale della popolazione e delle abitazioni, il primo del nuovo millennio (21 ottobre 2001), la maggior parte dei dati rilevati, ovvero quelli relativi alle persone residenti in famiglia (56.594.021), sono stati acquisiti tramite la lettura ottica e non attraverso il tradizionale *data entry*; inoltre, la codifica delle stringhe alfabetiche contenute sia nei Fogli di Famiglia che nei Fogli di Convivenza è stata effettuata attraverso software di codifica automatica in parte in *outsourcing* e in parte all'interno dell'Istat.

La codifica automatica dei testi ha rappresentato una delle principali innovazioni di processo del Censimento del 2001.

Si tratta di una attività svolta fino al 1991 dagli operatori dei comuni attraverso la consultazione dei manuali di classificazione delle singole variabili che comportava costi elevati sia in termini di tempo che di risorse umane impiegate, nonché una mancata standardizzazione del processo di attribuzione del codice.

Nei modelli di rilevazione predisposti per il Censimento del 2001 erano presenti cinque variabili testuali: Comune, Stato estero, Titolo di studio, Professione, Attività economica.

Le stringhe alfabetiche rilevate nel 2001 afferenti ai modelli acquisiti con tecniche Ocr/Icr[2] (Fogli di famiglia) sono state codificate in *outsourcing*, mentre quelle inserite nei Fogli di convivenza, nei Fogli di famiglia in lingua slovena e nei Fogli di famiglia integrativi,[3] registrate con il tradizionale *data entry*, sono state codificate all'interno dell'Istat tramite un sistema per la gestione della codifica che integrava il *software* ACTR[4] (*Automatic Coding by Text Recognition*) con il sistema informatico del censimento.

Test di codifica automatica precensuari basati su ACTR avevano, infatti, prodotto risultati incoraggianti rispetto ai *recall rate*[5] riferiti a descrizioni registrate con *data entry* ma non in relazione a testi acquisiti con lettura ottica, se non sottoposti ad una attenta videocorrezione.

La strategia del 2001 prevedeva che, per i Fogli di famiglia, venissero codificati solo i testi relativi alle variabili Comune, Stato estero e Titolo di studio mentre, per i Fogli di convivenza, era necessario processare anche le descrizioni di professioni e attività economiche. Infatti, dati i tempi lunghi che l'operazione di acquisizione e video-correzione degli stessi avrebbe implicato, nonché la complessità in termini di attribuzione dei codici alle descrizioni fornite dalla maggior parte dei cittadini per professione e attività economica (particolarmente articolate vista la natura delle variabili trattate) e la conseguente entità dei costi da sostenere, si era deciso di non procedere in prima battuta con l'assegnazione dei codici, bensì di prevedere a posteriori la codifica di un campione, stratificato per regione, dei testi acquisiti tramite lettura ottica.

Nel 2001 sono state predisposte a cura dell'Istat e consegnate alla ditta esterna prima dell'avvio del Censimento le basi informative (dizionari, uno o più per ciascuna variabile, cfr. Capitolo 3) e le linee guida sugli *step* da seguire nella fase di attribuzione dei codici, al fine di garantire, per quanto possibile, un livello di omogeneità accettabile nell'ambito di attività espletate da organi diversi.

In totale la Elsag, capogruppo del consorzio aggiudicatario della gara di appalto, ha provveduto a codificare oltre 58 milioni di stringhe. Le percentuali di attribuzione del codice, così come i livelli di accuratezza, sono stati verificati, per ogni singolo invio da parte della Elsag, da una seconda ditta esterna incaricata di certificare che ogni lotto di informazioni rispettasse, sia in termini quantitativi che qualitativi, tutti i parametri stabiliti in sede contrattuale, definiti coerentemente con i risultati ottenuti nelle sperimentazioni effettuate con le due indagini pilota (Tavola 1).

---

[2] Ocr: Optical character recognition. Icr: Intelligent character recognition.

[3] Si tratta di modelli forniti all'Istat dai comuni successivamente alle date stabilite per il ritiro dei modelli da parte del consorzio di ditte incaricate della lettura ottica.

[4] ACTR, studiato e commercializzato da Statistics Canada, è il sistema di codifica automatica generalizzato ed utilizzato in Istituto per la codifica delle risposte a testo libero fornite in molteplici indagini statistiche e che fanno riferimento a diverse classificazioni (S. Macchia, e al. 2007. Metodi e software per la codifica automatica e assistita dei dati).

[5] Percentuali di testi codificati automaticamente sul totale dei testi da codificare.

Tavola 1 - Parametri per l'attività di codifica in *outsourcing*

VARIABILIi	Livello minimo di assegnazione del codice	Livello minimo di accuratezza (a)
Comune	95%	99%
Stato estero	90%	98%
Titolo di studio	80%	98%

(a) Percentuale di codici corretti assegnati automaticamente sul totale dei codici assegnati.

L'attività di codifica delle variabili testuali rilevate nei Fogli di convivenza (mod.Istat CP.2) effettuata *in house* ha rappresentato l'obiettivo primario in funzione del quale all'interno dell'istituto è stato realizzato il sistema per la codifica automatica e *computer-assisted* delle stringhe alfabetiche.

Delle 674.174 descrizioni relative a Comune, Stato estero, Titolo di studio, Professione e Attività economica acquisite tramite il tradizionale *data entry* 566.060, ovvero l'83,8%, sono state codificate in *batch* mentre, nel 16,2% dei casi (108.114), è stato necessario l'intervento degli operatori manuali.

Tra i testi trattati nella fase *computer-assisted* solo nell'1,3% dei casi (0,2 per cento del totale delle descrizioni) non è stato possibile selezionare alcun codice.

Da un punto di vista quantitativo ACTR nel 2001 ha garantito livelli di attribuzione del codice in automatico prossimi a quelli conseguiti nelle precedenti applicazioni, con percentuali di codifica che si sono attestate da un minimo del 53.6%, rilevato in corrispondenza della variabile "attività economica", ad un massimo del 93.4% raggiunto per il "Comune".

Per quanto riguarda la qualità dell'informazione, dalle analisi condotte al termine dell'attività di codifica espletata all'interno dell'Istituto, l'accuratezza si è rivelata sempre coerente con i risultati ottenuti sui dati delle due indagini pilota precensuarie[6] (2001, Macchia, Mastroluca, Reale) e, in alcuni casi, superiore.

2. Le variabili testuali nel Censimento 2011 e le tecniche di rilevazione

Il 9 ottobre 2011 è la data di riferimento del 15° Censimento generale della popolazione e delle abitazioni, caratterizzato da innovazioni metodologiche progettate al fine di semplificare l'impatto organizzativo sulle amministrazioni pubbliche ed in particolare sui comuni, ampliare l'uso dei dati amministrativi, recuperare tempestività nella diffusione dei dati definitivi, ridurre il carico statistico sulle unità di rilevazione.[7.]

Quello del 2011, infatti, è stato un censimento assistito da liste; l'individuazione dei rispondenti è stata effettuata sulla base delle Liste Anagrafiche Comunali (LAC) e il completamento della rilevazione sul campo, a cura degli Uffici Comunali di Censimento, è stato supportato da liste ausiliare utili per l'individuazione di eventuali unità di rilevazione non presenti nelle LAC (ovvero predisposte a supporto del recupero della cosiddetta sottocopertura delle liste anagrafiche).[8]

A differenza del 2001, i questionari sono stati recapitati alle famiglie registrate nelle anagrafi comunali tramite spedizione postale. I rispondenti hanno avuto l'opportunità di compilare il questionario *online* o, in alternativa, di compilare il questionario cartaceo ricevuto a casa e restituirlo a un qualsiasi Ufficio Postale o presso uno dei centri di raccolta appositamente istituiti sul territorio comunale, con la possibilità di ottenere, sempre, la ricevuta di avvenuta consegna del modello di rilevazione compilato, indipendentemente dalla modalità di restituzione.

Il questionario elettronico, l'utilizzo delle LAC per la spedizione dei modelli di rilevazione alle famiglie e delle Liste Rnc (Rilevazione numeri civici), Lifa (Lista integrativa da fonti ausiliarie) e

[6] La prima e la seconda indagine pilota propedeutiche al censimento del 2001 sono state effettuate rispettivamente nel 1998 e nel 2000.

[7] A. Mancini "Il Censimento come investimento per il futuro delle statistiche demografiche e territoriali", 31° Convegno Nazionale ANUSCA, Riccione 17 novembre 2011.

[8] D. Zindato (a cura di) "15° Censimento generale della popolazione e delle abitazioni. Manuale della rilevazione", Istat - ottobre 2011.

Liac (Liste integrative autonome comunali) per il recupero della sottocopertura anagrafica rappresentano alcune delle più importanti innovazioni del Censimento del 2011.

A sostegno dell'attività di rilevazione è stato predisposto il Sistema di Gestione della Rilevazione (SGR), un supporto informatico indispensabile alla conduzione del censimento che ha fornito in tempo reale informazioni sull'andamento della rilevazione e che ha comportato, congiuntamente all'invio postale dei questionari e alla restituzione multicanale, una sensibile riduzione del numero di rilevatori necessario per le operazioni sul campo.

In occasione del Censimento del 2011, per la prima volta, alcune informazioni di carattere socio economico sono state rilevate su base campionaria. Questa nuova metodologia di rilevazione ha comportato la realizzazione di due tipi di modelli:

- uno "in forma ridotta" (Mod. Istat CP.1B), contenente le variabili necessarie per la produzione degli ipercubi che dovranno essere resi disponibili ad Eurostat ad un elevato dettaglio territoriale (NUTS3 - provinciale e LAU2 - comunale) e pochi altri quesiti utili anche nella fase di stima delle variabili inserite solo nei modelli "completi";
- uno "in forma completa" (Mod. Istat CP.1), contenente, oltre ai quesiti della forma "ridotta", tutte le altre variabili previste nel piano di rilevazione.

Nei comuni con almeno 20.000 residenti[9] o capoluoghi di provincia sono stati somministrati, in alternativa, i questionari in forma ridotta o in forma completa (33% di famiglie), mentre, nei comuni con classe di ampiezza demografica inferiore a 20.000 residenti, tutte le famiglie hanno ricevuto il questionario in forma completa.

Pertanto, le variabili inserite sia nei questionari in forma ridotta che in quelli in forma completa sono state rilevate in maniera esaustiva su tutta la popolazione residente; le variabili contenute solo nei questionari in forma completa, invece, sono state rilevate su tutte le famiglie residenti nei comuni con meno di 20.000 abitanti e su un campione di famiglie residenti (33%) nei comuni al di sopra di tale soglia demografica e nei capoluoghi di provincia.

Il modello CP.1B ha rappresentato una opportunità per ridurre significativamente il fastidio statistico su una parte di rispondenti. Caratterizzato da un numero contenuto di quesiti, assicura la disponibilità esaustiva di dati, oltre che demografici, anche socio economici della popolazione residente particolarmente rilevanti (grado di istruzione, stato occupazionale, spostamenti giornalieri all'interno del Comune o da Comune a Comune per motivi di studio o di lavoro).

Il questionario in forma completa approssima, come numero di quesiti, il Foglio di famiglia proposto al Censimento del 2001 e include, oltre alle variabili della forma breve, approfondimenti relativi all'istruzione, al lavoro e al pendolarismo e una nuova batteria di quesiti (non soggetti ad obbligo di risposta) atti a rilevare eventuali difficoltà (visive, uditive, deambulatorie, cognitive) che possono presentarsi nelle attività della vita quotidiana.

Tavola 2 - Struttura dei questionari CP.1B e CP.1

SEZIONI DEL QUESTIONARIO	Quesiti	
	Mod.Istat CP.1B (forma ridotta)	Mod.Istat CP.1 (forma completa)
Sezione I (famiglia e alloggio)	5	26
Sezione II (foglio individuale)	30	58
Totale (a)	35	84

(a) Riferito ad un solo foglio individuale (famiglia unipersonale).

Anche i modelli predisposti per questa tornata censuaria prevedevano alcune variabili testuali, ovvero: Comune, Stato estero e Titolo di studio.

[9] Per l'ampiezza demografica il dato è riferito al 31.12.2010.

Le descrizioni relative a Comune e Stato estero sono state richieste nei modelli in forma completa, in quelli in forma ridotta e nei Fogli di Convivenza (Mod.Istat CP.2); le descrizioni inerenti il titolo di studio solo nei modelli in forma completa e nei Fogli di Convivenza.

Nei questionari progettati per il 15° Censimento generale della popolazione e delle abitazioni non sono state inserite domande a risposta aperta su Professione e Attività economica. Si tratta, infatti, di variabili particolarmente complesse se rilevate nell'ambito di indagini sulle famiglie e che comportano elevati costi connessi all'acquisizione e alla codifica delle stringhe, motivo per cui i testi rilevati in occasione del Censimento del 2001 non sono stati né videocorretti né codificati.

Peraltro, il Regolamento Quadro dell'Unione Europea n.763/008 e il successivo Regolamento di attuazione della Commissione n.1201/2009[10] considera obbligatorie (*core topics*) le variabili *occupation* (Professione) fino al primo *digit* della classificazione ISCO 08 (i 10 grandi gruppi) e *industry* (Attività economica) fino a livello di sezione (NACE Rev.2). Proprio per questo, nei modelli CP.1 e CP.2 è stato inserito un quesito precodificato sull'attività lavorativa svolta, le cui modalità di risposta corrispondono ai 10 grandi gruppo della ISCO 08 e uno sui settori di attività economica che riporta, come risposte possibili, le 21 sezioni della NACE Rev.2.

Ciò premesso, le uniche variabili che devono essere sottoposte a un processo di codifica sono Comune (di nascita, di iscrizione anagrafica, di dimora abituale nel 2010, di dimora abituale nel 2006, di studio o di lavoro), Stato estero (di nascita, di cittadinanza, di cittadinanza precedente, di nascita dei genitori, di ultima residenza, di dimora abituale nel 2010, di dimora abituale nel 2006, di studio o di lavoro) e Titolo di studio. Nel caso del Comune, la codifica delle stringhe è subordinata all'indicazione della Provincia; per il Titolo di studio, l'attribuzione del codice deve essere effettuata in funzione delle modalità di risposta selezionate in corrispondenza dei quesiti precodificati relativi al grado di istruzione e alla durata del corso di studi.

La possibilità di restituzione multicanale, come già sottolineato, ha consentito di compilare il modello via internet oppure il questionario cartaceo ricevuto nel luogo di residenza.

Nel primo caso, la codifica delle variabili testuali è avvenuta contestualmente alla fase di compilazione del modello (Capitolo 5). Le stringhe riportate nei questionari cartacei vengono, altresì, acquisite in *outsourcing*, o tramite tecniche OCR/ICR (Mod.Istat CP.1 e Mod.Istat CP.1B) o tramite il tradizionale *data entry* (Mod.Istat Cp.2); lo stesso RTI incaricato della registrazione dei dati provvede poi anche alla codifica delle stringhe sulla base dei dizionari e delle linee guida predisposte dall'Istat (Capitolo 6).

Come già sottolineato, al Censimento del 2011 non sono state rilevate variabili testuali inerenti Professione e Attività economica, ma solo i relativi quesiti precodificati, peraltro già sperimentati nel Censimento di dieci anni fa.

Proprio in base all'esperienza del 2001, verificata all'epoca la difficoltà incontrata dagli utenti nel rispondere a questo tipo di quesiti, caratterizzati da modalità che riassumono in maniera non sempre chiara e immediata ampie categorie di professioni o di attività economiche, si è deciso di predisporre strumenti informatici ad *hoc* per supportare i rispondenti in fase di compilazione del questionario di censimento.

Si tratta del "navigatore delle Professioni" (cfr. Capitolo 7) e del "navigatore delle Attività Economiche" (cfr. Capitolo 8) attraverso cui, a partire dalla descrizione puntuale dell'attività lavorativa svolta o dell'Ateco a cui afferisce lo stabilimento/azienda in cui la persona lavora (o lavorava) o di cui è (o era) titolare, il rispondente ha potuto agevolmente individuare le modalità di risposta da selezionare.

I navigatori si aggiungono a tutte quelle innovazioni introdotte per il Censimento del 2011 con l'obiettivo di ridurre il carico statistico sui rispondenti e, contemporaneamente, incrementare la qualità del dato censuario.

[10] Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 on population and housing censuses, Official Journal L 218 , 13/08/2008 P. 0014 – 0020.
COMMISSION REGULATION (EC) No 1201/2009 of 30 November 2009 implementing Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses as regards the technical specifications of the topics and of their breakdowns.

3. La strategia di codifica e la logica per l'implementazione dei dizionari

Nell'ottica di automatizzare il processo di codifica delle variabili a testo libero, si è dovuto tener conto che la rilevazione era prevista, come già descritto nel Capitolo 2, secondo due diverse tecniche: compilazione del questionario su *Web* e compilazione del cartaceo con successiva acquisizione tramite lettura ottica. La funzione di codifica, per entrambe le tecniche di rilevazione, doveva essere tale da massimizzare i casi di successo, ossia di individuazione del codice, e di garantire la coerenza dei risultati.

Il primo fattore indispensabile per garantire la coerenza risiede nel fatto di condividere la stessa base informativa per entrambe le modalità di compilazione. È superfluo specificare che le classificazioni di riferimento per ciascuna variabile rilevata a testo libero costituiscano un punto di riferimento indubbiamente univoco, tuttavia è altrettanto evidente come utilizzare esclusivamente le dizioni ufficiali dei manuali delle classificazioni stesse non avrebbe assolutamente consentito di massimizzare i tassi di codifica.

L'assenza della figura dell'intervistatore nel censimento della popolazione, nonché l'eterogeneità dei rispondenti hanno quindi evidenziato la necessità di mettere a disposizione per la codifica basi informative ben più ampie rispetto a quanto contenuto nei manuali delle classificazioni.

Abitualmente, infatti, l'intervistatore, essendo appositamente formato sulle classificazioni di riferimento, sa orientarsi meglio del rispondente nello schema classificatorio cui il questionario elettronico accede qualora sia stata adottata una tecnica di rilevazione assistita da computer e, quando è richiesta la compilazione di un questionario cartaceo, tende ad orientare il rispondente nell'utilizzo di un linguaggio più appropriato.

Nel caso dell'autocompilazione, invece, il rispondente non può avere cognizione del fatto che la sua risposta dovrà essere ricondotta ad un codice secondo una classificazione ufficiale e risponde con il suo linguaggio, utilizzando vocaboli condizionati dal suo livello culturale, dalle conoscenze che ha a disposizione, nonché dall'area geografica in cui vive. Per di più nella rilevazione censuaria la numerosità dei rispondenti e la loro eterogeneità rendono questi fattori particolarmente rilevanti.

Per ciascuna delle classificazioni trattate, quindi, sono state predisposte apposite basi informative, i cosiddetti *dizionari*, che hanno per l'appunto la caratteristica di avvicinare il più possibile il linguaggio ufficiale delle classificazioni utilizzato nei manuali a quello con cui i rispondenti esprimono gli stessi concetti.

I passaggi logici per la costruzione dei dizionari sono concettualmente due: la rielaborazione dei manuali ufficiali e l'integrazione con dizioni alternative ma concettualmente analoghe rispetto a quelle utilizzate dai manuali stessi.

Nel dettaglio, la rielaborazione dei manuali delle classificazioni consiste essenzialmente:

- nel semplificare le descrizioni complesse: ci si riferisce ad esempio a quelle che associano allo stesso codice diversi concetti, mentre il rispondente tende ad esprimerne uno solo, quello che corrisponde alla sua realtà (se, per esempio, in una classificazione delle professioni, lo statistico e il matematico fossero associati allo stesso codice in un'unica dizione, il dizionario dovrà contenere, in corrispondenza di quel codice, due dizioni, una per lo statistico e una per il matematico, dal momento che, molto probabilmente, il rispondente dichiarerà di svolgere o l'una o l'altra professione);
- nel definire i sinonimi, intendendosi, ad esempio, gli elementi riconducibili ad un nome collettivo (legumi = fagioli, piselli, ecc..);
- nel rielaborare le classi aperte, che devono invece essere riempite di significato, in quanto nessun rispondente citerà "altro…." rispetto a quanto previsto dalla classificazione;
- nell'eliminare le clausole di esclusione, verificando che ciò che la classificazione esclude dall'associazione ad un certo codice, sia esplicitato in corrispondenza di un altro codice, in quanto, come per il caso delle classi aperte, nessun rispondente utilizzerà clausole di esclusione (ad esempio, se la classificazione delle Attività economiche associa ad un codice la dizione "coltivazione di cereali escluso il riso", nessun rispondente che coltiva cereali sentirà mai l'esigenza di specificare che non coltiva il riso);

- nell'effettuare integrazioni con materiale di riferimento, ad esempio descrizioni inerenti classificazioni associate a quella utilizzata (si pensi alla classificazione dei Prodotti rispetto a quella delle Attività economiche).

Le dizioni alternative a quelle utilizzate nei manuali si ricavano invece abitualmente da risposte empiriche precodificate fornite nell'ambito di precedenti indagini nelle quali è stato rilevato il fenomeno; chiameremo queste per convenzione *sinonimi*, anche se non nell'accezione semantica di questo termine. Si sottolinea infatti come quelli definiti sinonimi in questo contesto non debbano necessariamente essere sequenze di termini ortograficamente o sintatticamente corretti, purché però consentano l'individuazione di un codice univoco.

Coerentemente con quanto finora descritto, si è proceduto a costruire i dizionari per ciascuna delle classificazioni considerate (Provincia/Comune, Stato estero/Cittadinanza e Titolo di studio) e, come descritto nel Capitolo 4, la numerosità di descrizioni contenuta in ognuno di essi è risultata significativamente superiore rispetto a quella dei manuali delle classificazioni stesse. Si reputa interessante, in questo contesto, sottolineare alcuni aspetti che hanno influenzato la messa a punto dei dizionari per il censimento.

Il primo di questi riguarda il fatto che, tanto più i concetti associati a ciascun codice sono complessi, ossia possono dipendere da diversi fattori dalla cui interrelazione si evince il codice corrispondente, tanto più la definizione di *sinonimi* non può essere effettuata con un procedimento facilmente schematizzabile, generalizzabile e riproducibile automaticamente.

Per chiarire, si pensi alla classificazione dei Comuni e a quella dei Titoli di studio. Nella prima è possibile definire alcuni processi automatici nella generazione di *sinonimi* delle denominazioni ufficiali; ad esempio, nei comuni la cui denominazione contiene la parola 'Santo/a', si possono individuare tutti i modi in cui il rispondente può rappresentare questa parola (S., San, Santo, Sant', Santissimo, S.mo….); si può inoltre immaginare che il rispondente tenda ad utilizzare impropriamente la lettera Y e confonderla con la I e duplicare le denominazioni con entrambe queste lettere. Tali processi automatizzabili nella generazione di *sinonimi*, pur assorbendo parte non insignificante dell'attività di messa a punto dei dizionari, non sono senz'altro esaustivi per costruire la base informativa completa, in quanto diverse riflessioni puntuali devono essere effettuate per gestire ad esempio i comuni soppressi in quanto inclusi in altro Comune, oppure smembrati tra più comuni, quelli ceduti ad altro stato, ecc.,

Nel caso dei Titoli di studio, invece, la generazione dei *sinonimi* può essere difficilmente generalizzabile, in quanto per individuare un titolo di studio concorrono diversi aspetti, quali, oltre al nome che sintetizza la tematica prevalente del corso di studi (il liceo scientifico è così denominato ad indicare il fatto che predilige lo studio di materie scientifiche), la sua durata; inoltre spesso l'uso comune ha portato ad associare termini peculiari a diverse classi (ad esempio abitualmente il termine 'perito' è associato al diploma di istituto tecnico), oppure a descrivere il titolo di studio con sinonimi specifici dello stesso ("Economia e Commercio" – "Scienze economiche e commerciali"). Tutto ciò fa sì che la generazione dei sinonimi da includere nel dizionario abbia bisogno di riflessioni specifiche, spesso non generalizzabili e automatizzabili.

Da tutto ciò deriva che, mentre per le classificazioni dei Comuni e degli Stati esteri è stato possibile concordare alcuni principi generali in base ai quali procedere alla messa a punto dei dizionari, per il Titolo di studio è stata necessaria una riflessione puntuale che consentisse l'individuazione di *sinonimi* vicini al modo di esprimersi dei rispondenti e contemporaneamente non ambigui.

Un secondo fattore che ha avuto un impatto significativo nelle attività di messa a punto dei dizionari consiste nel fatto che, mentre per il Comune e lo Stato estero è stato possibile ripartire dalle basi informative predisposte per il precedente censimento e, a seguito del confronto con le nuove classificazioni, aggiornarle ed arricchirle, pur essendo necessario per alcuni sottoinsiemi ricostruirle ex novo seguendo però gli stessi processi logici, per il Titolo di studio è stato inevitabile ripartire da zero e recuperare soltanto una minima parte del lavoro già effettuato. La classificazione, infatti, ha subito una tale ristrutturazione che si è preferito costruire una nuova base informativa e recuperare da quella vecchia alcune dizioni afferenti soprattutto a titoli di studio non più esistenti e da ricondurre ai codici della nuova classificazione.

Il terzo aspetto considerato riguarda le due diverse tecniche di indagine utilizzate nel censimento: autocompilazione del questionario cartaceo e del questionare elettronico su *Web*.

3.1 Impatto della tecnica di rilevazione sulla messa a punto dei dizionari

È evidente che anche l'aspetto inerente la codifica di variabili testuali doveva rispondere al requisito di contenere al massimo l'effetto tecnica, consentendo quindi assoluta omogeneità nell'individuazione del codice.

Tuttavia, nella predisposizione dei dizionari da mettere in linea per ciascuna delle due tecniche, si è dovuto tener conto di diversi aspetti che si elencano di seguito in sintesi:

- con la rilevazione cartacea la fase di codifica avviene necessariamente in un momento successivo alla compilazione del questionario, quando non è più possibile l'interazione con il rispondente; viceversa con il questionari elettronico è il rispondente stesso a individuare il codice in fase di compilazione del questionario;
- come anticipato, con l'autocompilazione del questionario cartaceo non è possibile contenere la variabilità linguistica del rispondente, mentre con quello elettronico, per come è stata costruita l'applicazione, il rispondente, dopo aver digitato la sua risposta testuale, naviga nella base informativa messa a disposizione ed è quindi facilitato nell'individuazione del codice anche se non si realizza un *matching* esatto tra quanto digitato e le dizioni della base informativa;
- relativamente ai quesiti su Comune e Titolo di studio, con il questionario elettronico la navigazione nelle rispettive basi informative è circoscritta ad una porzione delle stesse, individuata dalla risposta ad un quesito precodificato. Infatti il testo digitato inerente la denominazione del Comune viene ricercato nella porzione di dizionario inerente la provincia esplicitata dal rispondente, così come il Titolo di studio viene ricercato soltanto tra quelli corrispondenti ad una delle 17 modalità di risposta del quesito precodificato; con la compilazione cartacea, invece, può non verificarsi la stessa cosa, in quanto si deve anche considerare il caso di mancata risposta ai quesiti precodificati oppure di incoerenza tra le risposte fornite ai quesiti precodificati e il testo digitato inerente la denominazione con cui effettuare il *mathing* testuale;
- un ulteriore aspetto da considerare nella predisposizione delle basi informative da mettere a disposizione per il questionario elettronico risiede nel fatto che da una parte la molteplicità delle descrizioni consente più probabilmente un *matching* esatto, dall'altra non è opportuno visualizzare un numero troppo elevato di descrizioni tra le quali il rispondente deve scegliere quella a lui più pertinente. Infatti, qualora il *matching* esatto non si realizzi, difficilmente il rispondente è portato a leggere un numero elevato di testi, tanto più se è necessario effettuare lo *scroll* della finestra visualizzata;
- infine, dal momento che, come descritto, i dizionari per la codifica per loro natura utilizzano descrizioni non ufficiali, questo aspetto deve essere chiaro per il rispondente, altrimenti quest'ultimo potrebbe arrivare alla conclusione che l'Istituto utilizza basi informative non corrette. Si pensi per esempio alle denominazioni dei comuni; è necessario evidenziare quale sia la denominazione ufficiale tra le diverse descrizioni associate allo stesso Comune che vengono visualizzate al rispondente a seguito della digitazione del testo di risposta al quesito. Questa problematica, ovviamente non si pone nel caso della compilazione del questionario cartaceo, per la quale il rispondente non è partecipe della fase di codifica.

Tutti questi aspetti hanno fatto sì che si rendesse opportuno differenziare i dizionari utilizzati per il questionario elettronico da quelli per la codifica in *batch*.

Come sarà dettagliato per ciascuna classificazione nel Capitolo 4, i dizionari utilizzati per la rilevazione su *Web* sono un sottoinsieme di quelli per la codifica in *batch* e sono corredati da una serie di informazioni ad uso del rispondente.

In sintesi i dizionari in linea per il *Web*:

- non contengono alcuni sinonimi che il rispondente può facilmente dedurre dalla visualizzazione delle finestre di *matching*;
- includono alcune etichette che identificano quale sia la denominazione ufficiale oppure specificano alcuni dettagli tipici di ciascuna classificazione (es. etichette di "Comune soppresso", "Stato soppresso", "Nome di territorio", ecc.).

4. Dalle classificazioni ai dizionari implementati

Come già detto nel Capitolo 3, per consentire l'individuazione di un codice a partire da una risposta a testo libero è stato necessario predisporre delle basi informative, ovvero dei dizionari che integrassero le dizioni ufficiali con un insieme di sinonimi in maniera tale da accostarsi il più possibile al linguaggio utilizzato dai rispondenti. Le basi informative utilizzate per la compilazione dei questionari elettronici e quelle per codificare i dati rilevati tramite il cartaceo sono logicamente coerenti ma non identiche. Mentre, infatti, con il questionario elettronico il rispondente è 'guidato nella compilazione' e può visualizzare il contenuto dei dizionari in apposite finestre, con il cartaceo non è previsto alcun controllo sulla correttezza e completezza dei testi rilevati nei quesiti da sottoporre a codifica. Per questo motivo, le basi informative da utilizzare per la codifica in *batch* sono più ricche di sinonimi e maggiormente dettagliate. L'ottica con cui si è lavorato per implementare i dizionari è stata quella, pertanto, di tenere in considerazione la variabilità dei testi usata dal rispondente in maniera tale di trovare un riscontro nei dizionari nella fase di codifica.

4.1 La struttura del codice statistico delle Unità amministrative

4.1.1 Alcune definizioni per la classificazione delle Province e dei Comuni

L'Istituto nazionale di statistica ha provveduto, a partire dagli anni sessanta, ad attribuire un codice statistico composto da sei cifre alle unità amministrative che esistevano al momento dell'osservazione (8.032 comuni e 92 province). Il codice statistico, strutturato in maniera gerarchica, prevede che le prime tre cifre identifichino la provincia di appartenenza del Comune, mentre le successive tre il Comune nell'ambito della stessa Provincia. Tale codice, pertanto, ha l'obiettivo di identificare le unità amministrative, indipendentemente dalla loro denominazione, in modo univoco nell'ambito territoriale di propria competenza.

I codici delle Province sono stati assegnati, inizialmente, secondo una numerazione progressiva a livello nazionale, seguendo il criterio dell'appartenenza alle regioni, nonché quello dell'ordinamento geografico, a partire da Nord-ovest verso Nord-est per passare poi verso il Centro, il Sud e le Isole maggiori. Per le province costituite dopo il 1966 è stato attribuito il numero successivo all'ultimo utilizzato. Tuttavia, al sistema della numerazione progressiva si affianca quello dell'ordinamento geografico qualora la classificazione riguardi più province neo costituite nell'ambito della stessa regione di appartenenza (l'esempio può essere ricondotto alla classificazione delle Province sarde istituite nel 2005). Una nota particolare riguarda, invece, le province cedute ad altra Nazione (tipicamente il caso della ex-Jugoslavia) che, classificate solo a partire dalla fine degli anni '90, sono così definite: 701 = Fiume, 702 = Pola, 703 = Zara.

La prima assegnazione dei codici dei Comuni è avvenuta seguendo l'ordine alfabetico di tali unità amministrative nell'ambito della Provincia di appartenenza.

Per i Comuni esistiti ed esistenti a partire da 1966, in relazione alle variazioni amministrative occorse, valgono le seguenti regole per la classificazione:

- un Comune che nel corso della sua storia non subisce variazioni di tipo amministrativo non cambia codice nel tempo;
- in caso di costituzione di un nuovo Comune, è attribuito il numero di codice successivo a quello dell'ultimo Comune della Provincia, indipendentemente dalla sua collocazione alfabetica;
- nel caso di soppressione di un Comune, il codice di quella unità amministrativa non viene ri-assegnato e cessa la sua validità a partire dalla data dell'evento;
- in caso di modifica della denominazione, il codice rimane invariato ed è assegnato alla denominazione attuale;
- in caso di costituzione di nuove Province, è attribuito un nuovo codice progressivo, secondo l'ordine alfabetico, ai Comuni della nuova Provincia. Di norma il codice dei Comuni della Provincia cedente rimane invariato, ma la regola è stata disattesa nel caso della costituzione delle Province di Oristano, Isernia e Pordenone, quando anche i Comuni delle Province cedenti di Cagliari, Campobasso e Udine sono stati rinumerati.

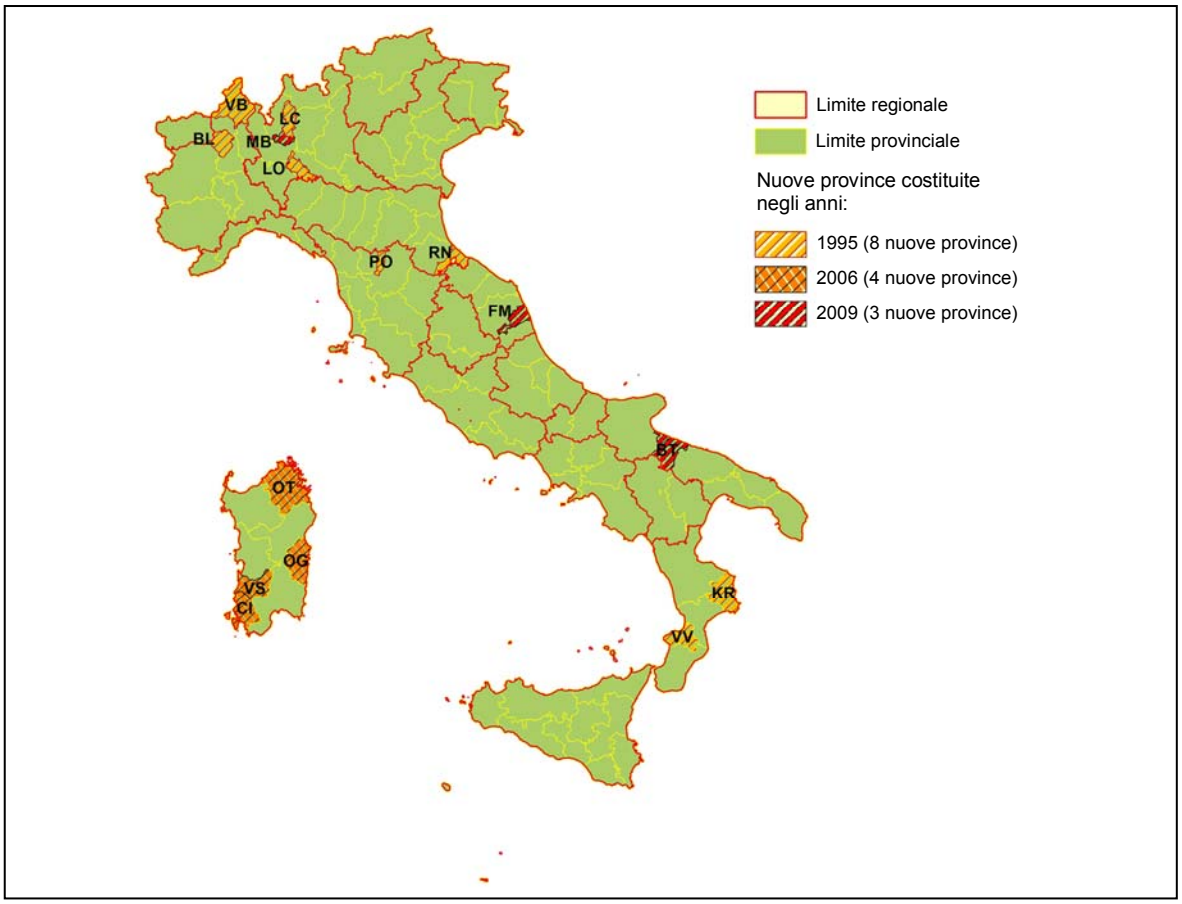
Per i Comuni soppressi e non ricostituiti prima del 1966, vale la seguente regola per la classificazione:

- è stato attribuito un codice progressivo, a partire da 801, ai Comuni elencati secondo l'ordine alfabetico nella Provincia di pertinenza territoriale al 1991, o di pertinenza dell'epoca, per quelli aggregati a Comuni successivamente ceduti alla ex-Jugoslavia.

Per i Comuni ceduti ad altra Nazione, vale la seguente regola per la classificazione:

- il codice è stato attribuito, con gli stessi criteri del punto precedente, a partire da 701, nell'ambito della Provincia di appartenenza al 1991, se esistente, o con codici delle Province di Fiume, Pola e Zara.

Figura 1 - Costituzione delle nuove Province



Fonte: Istat

4.1.2 Alcune definizioni per la classificazione delle località

Come esplicitato nel sottoparagrafo 4.1.5, il terzo dizionario messo a punto per il censimento e utilizzato per integrare le dizioni ufficiali con i sinonimi utilizzati dai rispondenti è quello delle località individuate per la realizzazione dei censimenti della popolazione e definite di seguito.

Nelle indagini statistiche la delimitazione dello spazio in cui si vuole indagare il fenomeno oggetto di studio dipende dall'obiettivo della rilevazione. L'indagine censuaria, per definizione, è una rilevazione totalitaria ossia viene effettuata con l'obiettivo di censire tutte le unità di rilevazione, ad esempio famiglie ed edifici per il censimento della popolazione, sul territorio nazionale. Ne segue che un primo elemento caratteristico, per la delimitazione dell'area interessata alla rilevazione, è definito sulla copertura completa dello spazio esaminato. Resta da definire in quale modo debba essere ripartita territorialmente questa copertura. Le basi territoriali (bt) rispondono esattamente a questa domanda, come dividere il territorio nazionale ai fini della rilevazione censuaria.

Una prima risposta viene dalla legge fondante della Repubblica Italiana. La Costituzione italiana nel titolo V, infatti, sancisce la divisione del suolo nazionale in regioni, province e comuni. Le basi territoriali sono quindi ripartite in questi livelli amministrativi. Ne segue anche una seconda importante considerazione, la copertura del territorio nazionale è completa e la gerarchia territoriale è necessariamente definita da oggetti geografici tra loro disgiunti e connessi. Tali proprietà fanno naturalmente riferimento ai confini con i quali sono definite le unità amministrative. All'interno del Comune, poi, la ripartizione del territorio deve rispondere al criterio di omogeneità spaziale ai fini di un maggior controllo della rilevazione statistica. In particolare il Comune viene ripartito in località abitate, divise in *centri* e *nuclei abitati*, in *località produttive* e nel resto del territorio comunale denominato *case sparse*. All'interno delle località sono delimitate aree di estensione territoriale minima, le sezioni di censimento. In alcune di queste le unità di rilevazione sono più numerose e vi è quindi la necessità di un maggior controllo della rilevazione in altre, di estensione maggiore, le unità di rilevazione sono in misura minore, come ad esempio nelle case sparse.

Lo scopo di questa operazione, infatti, è migliorare la fase di raccolta dei dati minimizzando gli errori dovuti alla mancata rilevazione o alla duplicazione delle unità censuarie. Non c'è dubbio, infatti, che una campagna di rilevazione così vasta ha la necessità di controllare capillarmente sul territorio lo svolgimento delle operazioni censuarie. Tuttavia, in questo contesto saranno definiti con maggior dettaglio solo alcuni aggregati, rimandando il lettore alla documentazione specifica per maggiori approfondimenti.

Le bt sono sancite nella legge 1228/1954, cui segue il Regolamento anagrafico e il relativo regolamento di attuazione e suoi aggiornamenti (D.P.R. 136/1958 e D.P.R. 223/1989). Il Regolamento anagrafico demanda all'Istat il compito di definire «le norme tecniche per l'esecuzione degli adempimenti dei comuni in materia topografica ed ecografica al fine di assicurarne uniformità e omogeneità d'applicazione» dove sono anche riportate le principali definizioni delle bt.

Pur nell'indubbia difficoltà di classificare situazioni territoriali così diverse anche tra le stesse località abitate, si pensi, ad esempio, alle differenze tra il centro abitato di Roma[11] e ad un piccolo centro di poche famiglie di un Comune montano, l'Istat divide le località abitate in *centri abitati* e *nuclei abitati*. La differenza sostanziale è la presenza di servizi per il cittadino come scuole, uffici pubblici, farmacie, negozi o simili. Con questa accezione, i centri abitati sono luoghi "auto contenuti" per chi vi risiede, ma anche per tutti i residenti che vivono in aree limitrofe senza servizi, in particolare, per i residenti delle case sparse e per quelle dei nuclei abitati dove chi vi abita si reca nei centri abitati anche per «le abituali esigenze della vita quotidiana». La definizione di *nucleo abitato*, invece, è particolarmente significativa anche sotto il profilo storico e dunque anche per l'associazione al codice comunale dei "modi di dire" caratteristici del territorio. La dispersione spaziale dei nuclei abitati delinea la struttura insediativa dell'Italia post-bellica fondata, generalmente, sulla vita della "masseria contadina" e sulle attività agricole. A questo fine è utile riportare la definizione di nucleo abitato del primo regolamento anagrafico del 1958 ancora vigente nel regolamento del 1992. «Il carattere di nucleo deve essere riconosciuto anche» […] «ai fabbricati di aziende agricole e zootecniche note nelle diverse regioni con varie denominazioni: *corte* (Lombardia), *casale* (campagna romana), *cassina o cascina* (Piemonte, Lombardia), *casanteria* (Romagna), *cussorgia e furriadroxius* (Sardegna), *villa* (Trentino), *colmello* (nel trevigiano), *maso* (Alto Adige), *borgo* (nel ferrarese) anche se costituiti da un solo edificio purché il numero di famiglie in esso abitanti non sia inferiore a cinque» (Istat, 1992).

Tuttavia, una visione più moderna delle località abitate mostra come, dal secondo dopoguerra fino ai nostri giorni anche per effetto del diverso modello di sviluppo economico, le località tendono ad accentrarsi e a crescere in estensione territoriale e demografica. In questo nuovo modello insediativo la soglia di 5 famiglie, indicata in passato, è un limite non più attuale per le odierne esigenze conoscitive della distribuzione territoriale della popolazione. L'Istat ha quindi aumentato a 15 il numero di famiglie residenti per consolidare un nuovo nucleo abitato. Ulteriori riflessioni sarebbero necessarie

[11] Il centro abitato di Roma, nel 2001, aveva una popolazione residente pari a 2.295.319, si veda la relativa pagina del *data warehouse* DaWinci sul sito ‹dawinci.istat.it›.

per giustificare questa decisione, ma non pertinenti in questo contesto. Per approfondire l'argomento si rimanda il lettore alla documentazione del progetto Census2010 (Istat, 2007).

Le *località produttive*[12] vengono introdotte con il Censimento del 2001. Lo scopo di questa operazione è delimitare sul territorio definito di case sparse le grandi concentrazioni di attività produttive, enuclearle dal territorio extraurbano dando una precisa collocazione spaziale e "battezzarle" con una denominazione propria.

Infine, il territorio comunale residuo, al netto delle località produttive e abitate, viene denominato di *case sparse*, ossia case «disseminate nel territorio comunale a distanza tale tra loro da non potere costituire nemmeno un nucleo abitato».

4.1.3 Le fonti utilizzate per aggiornare i dizionari delle Province e dei Comuni

Per il 15° Censimento della popolazione del 2011 le basi informative, ripartite nei tre Dizionari (sottoparagrafo 4.1.5), sono state aggiornate con il popolamento delle numerose variazioni amministrative (265 eventi in totale) intervenute dopo il Censimento della popolazione del 2001, nonché riorganizzate ed incrementate dall'inserimento di nuove tipologie di classificazione della "variabile Comune", anche in funzione delle necessità legate alla codifica interattiva per il questionario *on-line*.

L'elevato numero di descrizioni presenti nel sistema è stato alimentato dai risultati ottenuti dalla verifica periodica condotta presso le Regioni, di tutte le variazioni territoriali e amministrative verificatesi sul territorio nazionale e, a partire dal 2009, dalle attività di revisione delle Basi territoriali, strumento rispondente all'esigenza dei Comuni di predisporre i propri piani topografici per l'effettuazione dei Censimenti del 2011. Grande contributo è inoltre giunto da una continua ricerca delle fonti storiche e degli atti documentali attestanti le variazioni di cui sono stati oggetto le unità amministrative (comuni, province e regioni) dall'Unità d'Italia ad oggi.

L'aggiornamento delle nomenclature e delle codifiche contenute nei dizionari documentano, innanzitutto, come le variazioni amministrative occorse hanno portato il numero dei comuni italiani da 8.101 unità del Censimento della popolazione del 2001 a 8.092 unità in quello del 2011.

Il numero delle province italiane è, inoltre, passato da 103 a 110 unità, conseguentemente alla nascita delle sette nuove province (con il cambio di appartenenza di 228 comuni da province già esistenti a quelle di nuova istituzione) di: Olbia-Tempio, Ogliastra, Medio Campidano, Carbonia-Iglesias, Monza e della Brianza, Fermo e Barletta-Andria-Trani, nonché il passaggio di sette comuni dalla regione Marche (provincia di Pesaro e Urbino) a quella dell'Emilia-Romagna (provincia di Rimini).

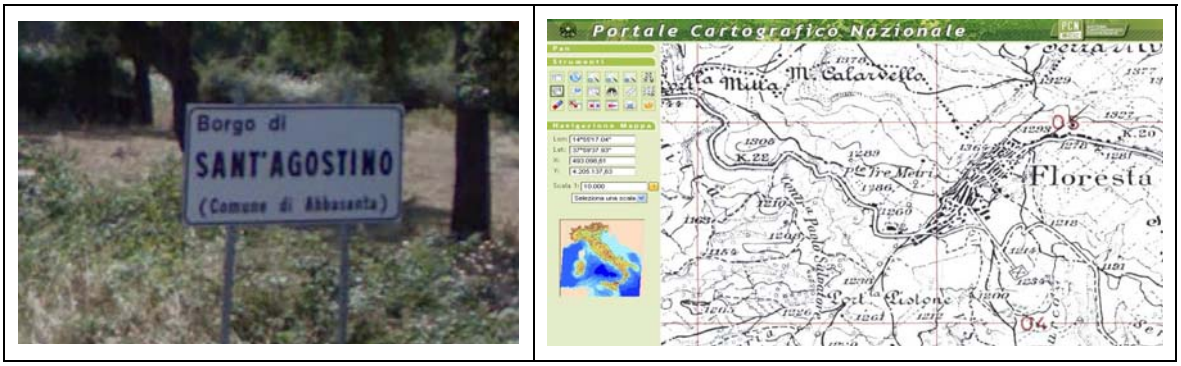
Inoltre, in seguito all'adozione del "Regulation (EC) n. 1059/2003 of the European Parliament and of the Council of 26 May 2003 on the establishment of a common classification of territorial units for statistics (NUTS)", è stato applicato l'uso della doppia dizione italiano/francese per la Regione e la provincia della Valle d'Aosta/Vallée d'Aoste e il ricorso al simbolo separatore "/" per i comuni con la denominazione bilingue della provincia di Bolzano/Bozen.

4.1.4 Le fonti utilizzate per aggiornare i dizionari delle località

La specifica attività di aggiornamento delle basi territoriali ha alimentato il terzo dizionario. Infatti, per ottenere il disegno delle bt ed i toponimi associati alle località, l'Istat produce una prima proposta di aggiornamento delle bt che viene poi rivista e validata dai comuni italiani. Qualora il Comune confermi l'esistenza di una località ma non riporti il toponimo di riferimento, l'Istat ha provveduto all'aggiornamento dei dati utilizzando delle fonti ancillari, quali ad esempio le informazioni provenienti dalle carte dell'Istituto geografico militare, disponibili in formato digitale sul sito www.atlanteitaliano.it, e le immagini di *Street View* di *Google Maps* (Figura 2).

[12] Area in ambito extraurbano non compresa nei centri o nuclei abitati nella quale siano presenti unità locali in numero superiore a 10, o il cui numero totale di addetti sia superiore a 200, contigue o vicine con interposte strade, piazze e simili, o comunque brevi soluzioni di continuità non superiori a 200 metri; la superficie minima deve corrispondere a 5 ettari.

Figura 2 - Fonti ancillari per la determinazione dei toponimi delle località (Street View e IGMI)



Fonte: Goolgle Maps Street View e IGMI

4.1.5 L'alimentazione dei dizionari delle Province, Comuni e località

La procedura di codifica è finalizzata ad attribuire alle risposte testuali, relative alla variabile Comune, un codice unico a 6 *digit*, nel quale i primi 3 rappresentano il codice provincia e gli ultimi 3 il codice Comune.

Per ottimizzare l'efficienza della procedura, i testi alla base della codifica sono stati organizzati in tre dizionari (Tavola 3) consultabili a cascata secondo un ordine definito: la ricerca del codice inizia all'interno del primo dizionario e si interrompe (non entrando quindi nel dizionario successivo) non appena viene individuato un codice univoco da attribuire alla risposta testuale da codificare.

1. **DIZIONARIO 1** (Diz1_comuni) – contenente le denominazioni ufficiali e i "modi di dire", ossia:
   - attuale denominazione degli 8092 Comuni italiani;
   - vecchia denominazione di alcuni Comuni;
   - denominazioni dei Comuni ricavate dalle risposte fornite nell'ambito di alcune indagini Istat: si tratta di "modi di dire" che si discostano dalla terminologia ufficiale per avvicinarsi al modo di esprimersi delle persone offrendo una maggiore possibilità di attribuire un codice univoco al testo da codificare;
   - denominazioni doppie, una in italiano e una in tedesco, per i Comuni in provincia di Bolzano. Per questi Comuni la denominazione ufficiale prevede due nomi, nelle lingue sopra citate, separati con uno *slash*; per ottimizzare la ricerca, alla dizione ufficiale sono state aggiunte, ovviamente con lo stesso codice, la dizione solo in italiano e la dizione solo in tedesco. Ad esempio:

BZ 021001	*Aldino/Aldein*	*(denominazione ufficiale)*
BZ 021001	Aldino	(denominazione italiana)
BZ 021001	Aldein	(denominazione tedesca)

   - denominazioni doppie per quei comuni con nomi composti, separati da un trattino o dalla congiunzione "e" (in modo analogo a quanto fatto per i comuni in provincia di Bolzano). Ad esempio:

RN 099001	*Bellaria-Igea Marina*	*(denominazione ufficiale)*
RN 099001	Bellaria	(denominazione sdoppiata)
RN 099001	Igea Marina	(denominazione sdoppiata)

2. **DIZIONARIO 2** (Diz2_comuni) – contenente Comuni non più esistenti in quanto aggregati ad uno o più Comuni oppure passati ad uno Stato estero, Comuni che hanno cambiato Provincia e rioni. In maggior dettaglio le informazioni presenti sono:
   - denominazioni di Comuni soppressi ed aggregati ad un altro Comune;
   - denominazioni di Comuni soppressi ed aggregati a più Comuni;
   - denominazioni di Comuni passati ad un altro Stato (si tratta di comuni ubicati al confine, ceduti generalmente allo stato della ex Jugoslavia);

- Comuni che hanno cambiato Provincia in seguito all'istituzione di nuove province o al cambiamento di alcune sigle: per ottimizzare la codifica è stato previsto di inserire anche i comuni con la vecchia sigla della provincia fermo restando il codice numerico corrispondente alla provincia attuale;
- denominazioni dei rioni, ossia dei quartieri che compongono le grandi città.

3. **DIZIONARIO 3** (Diz3_comuni) – contenente le denominazioni delle località, con una serie di varianti, analoghe a quelle riportate nei dizionari 1 e 2.

Il criterio adottato per stabilire l'ordine di consultazione dei dizionari è basato sul principio di risposta "più probabile": il primo dizionario contiene le risposte che, più probabilmente, i rispondenti indicheranno ad una domanda concernente il Comune, *ad es.* "*Comune di residenza*", fornendo il nome ufficiale o quello precedente, qualora sia stato modificato, o il modo di chiamare il Comune. Il secondo dizionario contiene sostanzialmente Comuni soppressi che hanno quindi una probabilità minore di essere citati rispetto a quelli appartenenti al primo gruppo. Ad esempio nell'indicare il "*Comune di nascita*" un rispondente potrebbe riportare la dizione del periodo in cui il Comune era a se stante piuttosto che indicare il nome del nuovo Comune di appartenenza. Infine il terzo dizionario contiene le località che il rispondente potrebbe indicare invece di fornire la risposta corretta del Comune.

I tre dizionari condividono tra loro la stessa struttura che però è un po' diversa tra i dizionari utilizzati per codificare i dati con il cartaceo e quelli messi in linea per il questionario elettronico, in quanto per il *Web* erano necessarie alcune informazioni che sarebbero state ridondanti per effettuare la codifica in *batch*.

Tavola 3 - Numerosità dei tre dizionari

DIZIONARIO	Dimensione/Nr *record*
Dizionario1	12.651
Dizionario2	4.294
Dizionario3	56.419

A ciascuna denominazione dei dizionari è stato fatto corrispondere un *flag* rappresentativo di un'informazione, ininfluente ai fini della codifica, ma che può essere invece molto utile per l'aggiornamento dei dizionari stessi. Tale *flag*, infatti, sta a significare, ad esempio, se trattasi di attuale Comune oppure di Comune soppresso, se la dizione non è quella ufficiale, ma frutto di trasformazioni effettuate sulla stessa al fine di migliorare il risultato di codifica, eccetera.

Nella tavola di seguito riportata è esplicitato il significato dei codici associati a questo *flag*.

Tavola 4 - Codici esplicativi delle dizioni dei dizionari

DIZIONARIO	*Flag* Identific.	Occorrenze	Descrizione
	(1__)	8.092	Denominazioni ufficiali (correnti all'attuale data)
	(1a__)	144	Denominazioni ufficiali (correnti all'attuale data) con l'accento finale apostrofato
	(10_)	32	Denominazioni ufficiali con carattere J oppure j
	(5__)	14	Denominazioni ufficiali con carattere Y oppure y
	(105)	1	Denominazioni ufficiali con carattere J o j e con carattere Y oppure y
	(11_)	226	Comuni di BOLZANO con dicitura solo Italiana o Tedesca) (*)
Dizionario 1	(110)	3	Comuni di BOLZANO sdoppiati e con carattere J oppure j
	(3__)	2.460	Comuni con cambio denominazione
	(30_)	15	Comuni con cambio denominazione e con carattere J oppure j
	(35_)	3	Comuni con cambio denominazione e con carattere Y oppure y
	(37_)	17	Comuni con cambio denominazione e con carattere "-"
	(6__)	1.608	Modi di dire (empiriche rilevate in una precedente indagine)
	(7__)	36	Nome di Comune con carattere "-"

Tavola 4 segue **- Codici esplicativi delle dizioni dei dizionari**

DIZIONARIO	*Flag* Identific.	Occorrenze	Descrizione
	(2__)	1.769	Ex comuni che sono stati aggregati a un altro Comune
	(20_)	7	Come [2] e con carattere J oppure j
	(25_)	1	Come [2] e con carattere Y oppure y
	(27_)	240	Come [2] e con carattere "-"
	(270)	3	Come [2] e con carattere "-"e con carattere J oppure j
	(275)	1	Come [2] e con carattere "-"e con carattere Y oppure y
	(28_)	281	Come [2] e con che hanno cambiato sigla provincia
	(4__)	134	Comuni che attualmente sono in territorio estero
	(42__)	81	Come [2] ma che sono in territorio estero
	(43__)	43	Come [3] ma che sono in territorio estero
	(47_)	2	Come [4] e con carattere "-"
	(48_)	5	Come [37] ma che sono in territorio estero
Dizionario 2	(8__)	914	Comuni che hanno cambiato sigla provincia
	(87_)	7	Come [8] e con carattere "-"
	(38_)	233	Come [3] ma che hanno cambiato sigla provincia
	(68_)	71	Modi di dire (da una precedente indagine) ma che hanno cambiato sigla
	(108)	2	Denominazioni ufficiali con carattere Y o y e cambiamento sigla provincia
	(158)	2	Denominazioni ufficiali con carattere J o j e cambiamento sigla provincia
	(9__)	137	Ex comuni che sono stati smembrati ed aggregati a più comuni
	(90_)	2	Come [9] e con carattere J oppure j
	(97_)	3	Come [9] e con carattere "-"
	(98_)	4	Come [9] ma che hanno cambiato sigla provincia
	(99_)	21	Come [9] ma con cambio denominazione
	(b__)	319	*Rioni* di comuni
	(b8_)	12	*Rioni* di comuni che hanno cambiato sigla provincia
	(a__)	51.720	*Località* di alcuni comuni
	(a_8)	4.335	Località i cui comuni hanno cambiato sigla provincia
	(a0_)	107	Località con il carattere J oppure j
	(a05)	2	Località con il carattere J o j e carattere Y o y
Dizionario 3	(a08)	14	Località con il carattere J o j e cambio sigla provincia
	(a5_)	115	Località con il carattere Y oppure y
	(a7_)	110	Località con denominazioni con carattere "-"
	(a78)	16	Località con denominazioni con carattere "-" e cambio sigla provincia

4.1.6 Il dizionario per la codifica con il questionario Web

È necessario premettere che per la codifica con il questionario Web non è stato messo in linea il dizionario 3 con le località, in quanto si è ritenuto che la possibilità di visualizzare l'elenco dei nomi dei comuni desse sufficienti garanzie per l'individuazione del Comune di competenza.

Inoltre i dizionari 1 e 2 sono stati messi in linea come un unicum e corredati da più colonne rispetto a quelle utilizzate per la codifica in *batch.*

In particolare, la loro struttura è la seguente:

- **Colonna A**: sigla provincia (non sempre è la sigla dell'attuale provincia in cui si trova il Comune, ossia di quella che corrisponde ai primi tre *digit* del codice Comune della colonna C; lo è sempre per i *record* con *flag* 1);
- **Colonna B** codice provincia corrispondente alla sigla della colonna A;
- **Colonna C** codice Comune (6 *digit*, dei quali i primi tre sono i codici dell'attuale provincia);
- **Colonna D** *Flag* (come da Tavola 4);
- **Colonna E** Denominazione Comune;

- **Colonna F** sigla della provincia corrispondente alla sigla provincia del *record* con *flag* 1, ossia ai primi tre *digit* del codice Comune della colonna C;
- **Colonna G** Denominazione ufficiale del Comune e altro;
- **Colonna H** Codice provincia corrispondente alla sigla presente nella colonna F (è un di più, in quanto tale codice coincide con i primi tre *digit* del codice di col. C).

Rispetto al totale delle descrizioni contenute nei dizionari, non sono inoltre stati utilizzati i nomi con accento finale apostrofato, identificati con il *flag* = 1a.

È stata infine introdotta una ulteriore differenzazione tra l'elenco di nomi da mettere a disposizione per il quesito sul Comune di nascita rispetto a quelli sugli altri comuni (residenza, lavoro…). Infatti, mentre per il quesito sul Comune di nascita dovevano essere utilizzati ai fini del *matching* tutti i *record*, in quanto le risposte si riferivano alla situazione dei comuni corrispondente all'anno di nascita, per i quesiti su Comune di residenza e lavoro ci si riferiva ad una realtà più recente, quindi non sono stati considerati i *record* con *flag* 4, 42, 43, 47, 48 (attualmente Stato estero) e 9, 90, 97, 98 e 99 (Comuni non più esistenti).

Da ultimo, per il questionario *online*, è stati utilizzato per il *display* anche il contenuto delle colonne **F** e **G,** in aggiunta al contenuto della colonna E; in queste colonne, infatti, è riportata la sigla dell'attuale provincia e la denominazione ufficiale del Comune. È importante che sia visualizzato il *display* della sigla dell'attuale provincia soprattutto per i Comuni che hanno cambiato provincia, ma che sono stati ricercati nella vecchia provincia, in quanto il rispondente aveva indicato quest'ultima nel precedente quesito; il *display* della denominazione ufficiale è invece importante per i comuni individuati grazie al *matching* con una dizione 'sporca', diversa da quella ufficiale, oppure con la denominazione di vecchi comuni non più esistenti.

4.1.7 Il dizionario per la codifica in batch di quanto rilevato con il questionario cartaceo

Come anticipato, per la codifica in *batch* sono stati messi a disposizione tutti e tre i dizionari nell'ottica di massimizzare i tassi di codifica.

La struttura dei tre dizionari corrisponde a quella riportata nella Tavola 5.

Tavola 5 - Struttura dei dizionari utilizzati nel questionario cartaceo

CAMPO/COLONNA	Dimensione	Descrizione
A	2	Sigla provincia (da utilizzarsi come filtro in funzione della sigla riportata nel corrispondente quesito del questionario PROV) (attenzione: non sempre è la sigla dell'attuale provincia in cui si trova il Comune, ossia di quella che corrisponde ai primi tre *digit* del codice Comune della colonna B; lo è sempre per i *record* con *flag* 1_ a colonna C)
B	6	Codice da attribuire in caso di match Si tratta di: Codice Provincia (attuale) + Codice Comune (attuale)
C	Variabile (max 5)	Identificativo relativo alla "tipologia" del Comune e alle trasformazioni effettuate sulle denominazioni ufficiali, non utilizzato ai fini della codifica (*flag* come da Tavola 4)
D	(variabile)	Denominazione Comune (testo su cui effettuare il match testuale)

4.2 Titolo di studio

4.2.1 La classificazione del Titolo di studio

La sola classificazione generale dei Titoli di studio attualmente disponibile a fini statistici è, quella prodotta dall'Istat nel 2003, frutto di un processo di aggiornamento e affinamento avviato a seguito della prima esperienza realizzata in occasione del Censimento della popolazione del 2001.

Si tratta di una classificazione costruita principalmente con fonti statistiche con l'obiettivo di ricostruire, in maniera il più possibile estensiva, l'insieme dei Titoli di studio emessi dal sistema scolastico e universitario del nostro Paese e potenzialmente in possesso della popolazione italiana.

In occasione del Censimento 2011 tale classificazione è stata aggiornata e implementata in modo da tenere conto delle trasformazioni avvenute nel sistema educativo nel decennio 2001-2011 e,

quindi, dei nuovi titoli rilasciati dal sistema scolastico e universitario in quel periodo. Anche la nuova versione della classificazione, pur contenendo alcune differenze rispetto a quella utilizzata nel 2001, mantiene una struttura di tipo gerarchico (che va dal titolo di livello iniziale a quello di livello più elevato) in coerenza con la Classificazione internazionale ISCED 97,[13] in modo da assicurare la comparazione a livello internazionale dei singoli titoli. Ogni Titolo di studio è caratterizzato da un codice a 8 *digit*:

1. i primi due *digit* indicano la posizione del titolo (livello) rispetto alla scala gerarchica;
2. il terzo, quarto e quinto *digit* indicano la tipologia della scuola (liceo, istituto tecnico, ecc.) o il gruppo disciplinare del corso accademico (scientifico, medico, ecc.);
3. gli ultimi tre *digit* indicano lo specifico indirizzo disciplinare del corso scolastico/universitario (codice progressivo) all'interno del tipo scuola/gruppo accademico.

A differenza dei primi due, i codici dal terzo all'ottavo non seguono alcun ordinamento gerarchico. La concatenazione dei tre codici definisce il complessivo codice del Titolo di studio.

Per quanto riguarda la scuola secondaria di II grado, in particolare, l'elenco dei titoli è stato implementato con le informazioni acquisite tramite il Ministero dell'Istruzione in merito ai titoli conseguiti al termine dei percorsi "sperimentali", per i quali non è tuttavia disponibile una ricognizione completa e ufficiale.

Le novità principali hanno riguardato i titoli rilasciati dal sistema terziario (università e corsi accademici dell'istruzione artistica e musicale). Per quanto riguarda l'università, pur lasciando inalterata l'architettura del sistema introdotto con il decreto 509/99, che aveva istituito il cosiddetto sistema del 3+2 (laurea di I livello e laurea specialistica di II livello), il D.M. 270/2004 ha determinato una revisione delle classi delle lauree e delle lauree specialistiche (passate rispettivamente da 42 a 43 e da 104 a 94 e da allora definite "magistrali") e ha introdotto una nuova classe di laurea magistrale a ciclo unico in Giurisprudenza.

Al fine di rilevare in modo omogeneo sia le lauree del nuovo che del vecchio ordinamento (nel quale le classi non erano previste) nel quesito "aperto" si è deciso di richiedere l'indicazione dello specifico corso di laurea. Pertanto, rispetto alla Classificazione Istat dei Titoli di studio 2003, in quella utilizzata per il Censimento 2011 i *digit* sesto, settimo e ottavo si riferiscono al singolo corso di laurea e non alla "classe". Ciò ha comportato una impegnativa opera di riclassificazione dei corsi del nuovo ordinamento, in particolare di quelli che, pur avendo la stessa denominazione, afferiscono a classi diverse. La maggior parte di questi "doppioni", tuttavia, sono stati classificati senza problemi all'interno di uno stesso gruppo disciplinare, mentre in alcuni, pochi casi è stato necessario operare una forzatura in quanto afferivano a gruppi diversi. Per attribuire il codice in questi casi è stato utilizzato il criterio della "prevalenza" (numero complessivo di laureati fino al 2011 e diffusione sul territorio).

La classificazione, infine, è stata implementata con l'aggiunta dei corsi del nuovo sistema di Alta Formazione Artistica, Musicale e Coreutica (AFAM) istituiti dalla legge 58/1999 e attivati a partire dal 2004.

Il nuovo sistema di livello terziario si articola su tre cicli di studio in conformità con il sistema universitario. Rimangono comunque ancora attivi i corsi accademici del vecchio ordinamento, i quali, appartenendo ad un diverso "livello gerarchico", mantengono una codifica distinta.

4.2.2 La messa a punto del dizionario

La procedura è finalizzata alla codifica delle risposte testuali rilevate al quesito Q.5.5.[14] Tale quesito è preceduto da un quesito precodificato, il Q.5.3, in cui sono previste diciassette modalità di risposta relative ad altrettanti livelli di istruzione (Tavola 7).

Il dizionario, predisposto in formato Excel, è stato strutturato come nella seguente tavola 6.

[13] International Standard Classification of Education, 1997, Unesco - Oecd - Eurostat.

[14] La numerazione dei quesiti si riferisce al modello CP.1; i corrispondenti quesiti relativi al modello CP.2 sono il 4.3, 4.4 e 4.5.

Tavola 6 - Struttura del dizionario

COLONNA	Dimensione	Descrizione
A	3	Filtro corrispondente alla biffatura (Quesito Precodificato Q 5.3)
B	8	Codice da attribuire in caso di *match* I codici non hanno tutti la stessa lunghezza: quelli corrispondenti alle biffature del quesito precodificato da 05 a 12, 14 e 16 → codice a 5 *digit* quelli corrispondenti alle biffature del quesito precodificato 13, 15 e 17→ codice a 8 *digit*
C	1	*Flag* corrispondente alle dizioni visualizzate dei titoli di studio nel questionario *Web*
D	(variabile)	Descrizioni corrispondenti ai vari Titoli di studio (testo su cui effettuare il *match* testuale)

Nel dettaglio, si specifica che la **Colonna A** contiene valori da utilizzare come filtro per l'individuazione del codice da attribuire. Tali valori corrispondono alle possibili modalità di risposta del quesito precodificato (Q 5.3) del questionario. Tale colonna può essere anche vuota come verrà meglio specificato nei sottoparagrafi successivi.

La **Colonna B** contiene sia i codici della Classificazione ufficiale che alcuni codici fittizi, quest'ultimi utilizzati soltanto nella codifica in *batch* delle risposte fornite con la rilevazione cartacea. I codici hanno diversa lunghezza a seconda che afferiscano ai titoli di scuola secondaria superiore o alle accademie (lunghezza codice 5 *digit)*, oppure a quelli rilasciati dalle università (lunghezza codice 8 *digit*).

L'informazione, invece, riportata nella **Colonna C** è ininfluente per la codifica in *batch*, in quanto è funzionale esclusivamente alla codifica in corso di intervista con il questionario *Web*; indica, infatti, il sottoinsieme di testi con i quali effettuare il *matching* e da visualizzare a seguito di questo.

Nella **Colonna D** viene riportato l'elenco dei Titoli di studio. Questo elenco è stato costruito in modo di avvicinare quanto contemplato dalla Classificazione ufficiale al linguaggio utilizzato dai rispondenti. Tale elenco comprende quindi:

- i Titoli di studio previsti dalla Classificazione ufficiale;
- i diversi modi di esprimere i Titoli di studio, ricavati da precedenti Indagini nell'ambito delle quali è stato rilevato il fenomeno.

Si sottolinea quindi che queste ultime descrizioni, non essendo quelle *'previste dai manuali'*, utilizzano una terminologia variegata, a volte anche bizzarra, con la quale comunque il rispondente ha fornito un'informazione utilizzabile a fini elaborativi.

Per esempio chi ha risposto di possedere la "*qualifica artigianale*" (terminologia del rispondente) è stato ricondotto al filtro 061 (Tavola 7) ed al codice 30102, corrispondente al *Diploma di qualifica di istituto professionale per l'industria e l'artigianato* (corso di studi della durata di 2-3 anni) ovvero al Titolo di studio proprio della Classificazione ufficiale.

4.2.3 Il dizionario per la codifica con il questionario Web

Nei questionari di Censimento predisposti per il 2011, il quesito che rileva la variabile testuale "*Titolo di studio*", dove viene richiesto di specificare il titolo più elevato conseguito, è preceduto da un quesito precodificato (Q.5.3) in cui il rispondente individua con una biffatura il livello di istruzione e, solo per alcune modalità di risposta, da un altro relativo alla durata del corso di studi (Q.5.4).

I quesiti precodificati nella compilazione del questionario *online* rappresentano campi obbligatori; in questo modo è stato possibile restringere la ricerca testuale solo all'interno del sottoinsieme di dizionario corrispondente alle modalità selezionate. Per ottenere questo risultato, il dizionario è stato corredato da un filtro i cui valori sono stati stabiliti in funzione delle risposte fornite nei quesiti precodificati come mostra la tavola che segue.

Tavola 7 - Corrispondenza tra modalità di risposta del quesito precodificato e dizionario

FILTRO	Modalità Q.5.3	Descrizione titolo di studio
010	01	Nessun titolo di studio e non sa leggere o scriver
020	02	Nessun titolo di studio ma sa leggere e scrivere
030	03	Licenza di scuola elementare (o valutazione finale equivalente)
040	04	Licenza di scuola media (o avviamento professionale)
050	05	Compimento inferiore/medio di Conservatorio musicale o di Accademia Nazionale di Danza (2-3-anni)
061	06-e 5.4 = 1	Diploma di istituto professionale (2-3 anni)
062	06-e 5.4 = 2	Diploma di istituto professionale (4-5 anni)
071	07-e 5.4 = 1	Diploma di scuola magistrale (2-3 anni)
072	07-e 5.4 = 2	Diploma di scuola magistrale (4-5 anni)
081	08-e 5.4 = 1	Diploma di Istituto d'arte (2-3 anni)
082	08-e 5.4 = 2	Diploma di Istituto d'arte (4-5 anni)
090	09	Diploma di Istituto tecnico
100	10	Diploma di Istituto magistrale
110	11	Diploma di liceo (classico, scientifico, ecc.)
120	12	Diploma di Accademia di Belle arti, Danza, Arte drammatica, ISIA, ecc. Conservatorio (vecchio ordinamento)s
130	13	Diploma universitario (2-3 anni) del vecchio ordinamento (incluse le scuole dirette a fini speciali o parauniversitarie)
140	14	Diploma accademico di Alta Formazione Artistica, Musicale, e Coreutica (A.F.A.M.) (di I livello)
150	15	Laurea triennale (di I livello) del nuovo ordinamento
160	16	Diploma accademico di Alta Formazione Artistica, Musicale, e Coreutica (A.F.A.M.) (di II livello)
170	17	Laurea (4-6 anni) del vecchio ordinamento, laurea specialistica o magistrale a ciclo unico del nuovo ordinamento, laurea biennale specialistica (di II livello) del nuovo ordinamento

Il quesito, come si vede, prevede ben 17 modalità di risposta (dodici delle quali richiedono un processo di codifica). In relazione al questionario elettronico, ovvero durante la codifica *online*, si è stabilito che:

- per le modalità di risposta da 05 a 12, 14 e 16 sono state previste apposite tendine entro le quali il rispondente può individuare il proprio Titolo di studio;
- in particolare per le modalità 06, 07 e 08, le tendine sono state vincolate alla risposta fornita sulla durata del corso di studi (Q.5.4);
- per le modalità di risposta 13, 15 e 17 è stata predisposta, invece, una ricerca testuale su apposite tabelle (una per ciascuna modalità di risposta) nell'ambito delle quali il rispondente può individuare il proprio Titolo di studio compatibile con la risposta fornita al quesito precodificato.

Per i Titoli di studio universitari (modalità 13, 15 e 17), nella fase di codifica ci si è avvalsi di un algoritmo per il *matching* testuale che confronta la "descrizione risposta" con le "descrizione del dizionario" sulla base di funzioni di similarità tra testi (Capitolo 5). Avvalendosi dell'esperienza maturata nell'implementazione di contesti di codifica automatica con ACTR, sono stati forniti degli elenchi di parole chiamate *stopword*, ovvero parole che sono state ritenute ininfluenti ai fini del *match* e che quindi possono essere tolte dal testo della "descrizione risposta".

Tra le parole inutili valide per tutte e tre le modalità di risposta, sono state inserite le preposizioni semplici ed articolate, gli articoli, e le parole come *diploma, universitario, corso* ecc.

Per la modalità di risposta 13 (Diplomi universitari del vecchio ordinamento, Scuole dirette a fini speciali o parauniversitarie) sono state considerate inutili parole del tipo: *parauniversitario, minilaurea* ecc. Per la modalità di risposta 15 (Lauree triennali del nuovo ordinamento) sono state considerate inutili parole del tipo: *primo, livello, triennale* ecc.

Per la modalità 17 (Laurea del vecchio ordinamento, Laurea specialistica o magistrale a ciclo unico del nuovo ordinamento, Laurea biennale specialistica di II livello del nuovo ordinamento) non è stato possibile individuare, invece, alcuna parola inutile a parte gli articoli e le preposizioni, così come esplicitato sopra, in quanto in essa sono accorpati più titoli di studio afferenti sia al nuovo che al vecchio ordinamento più le lauree a ciclo unico.

Il Dizionario messo in linea per il questionario elettronico, quindi, ha la stessa struttura descritta nel sottoparagrafo 4.2.2, con la peculiarità che il campo filtro della colonna A corrisponde esclusivamente alle biffature 13, 15 e 17.

Si specifica inoltre relativamente al *flag* = 1, posto a Col C (riguarda solo i *record* che debbono essere visualizzati nella fase di *matching*) che:

1. nella modalità 13 hanno *flag* = 1 esclusivamente le denominazioni dei *Diplomi universitari del vecchio ordinamento* senza l'aggiunta di alcuna specifica (in totale 272 *record*);
2. nella modalità 15 è stato necessario apporre il *flag* = 1 a tutti i *record* (in totale 1773 *record*);
3. nella modalità 17 per ogni Titolo di studio è stato necessario apporre il *flag* = 1 in corrispondenza sia a *Laurea proveniente dal vecchio ordinamento* che a *Laurea specialistica/magistrale del nuovo ordinamento*, in quanto non è possibile evincere dalla biffatura il Titolo di studio posseduto dal rispondente (in totale 4511 *record*). Chiarendo meglio con un esempio (Tavola 8), se l'intervistato dichiara genericamente di possedere la '*Laurea in matematica*' gli apparirà un elenco (vedi sotto) nel quale avrà la possibilità di individuare in maniera dettagliata il corso di studi effettivamente frequentato.

Tavola 8 - esempio di descrizioni afferenti alla modalità di risposta 17

FILTRO	Codice	F	Titolo di studio
170	72001001	1	Laurea Matematica (vecchio ordinamento)
170	74001056	1	Matematica (Laurea specialistica nuovo ordinamento)
170	74001056	1	Laurea magistrale Matematica

L'insieme delle dizioni utilizzate per la codifica *online* dei Titoli di studio universitari è costituito, pertanto, da un sottoinsieme del dizionario predisposto per il *matching* testuale ovvero solo dai *record* che presentano a colonna C il *flag* = 1 (in totale 6556 *record*).

4.2.4 Il dizionario per la codifica in batch di quanto rilevato con il questionario cartaceo

Per la codifica *batch* sono state utilizzate, invece, tutte le descrizioni del dizionario, in quanto con il questionario cartaceo era richiesta una risposta testuale anche per le modalità da 05 a 12, 14 e 16, a differenza della codifica *online* per la quale erano state implementate le apposite 'tendine'.

È stato, inoltre, necessario predisporre l'ambiente informativo in modo da gestire una serie di casistiche, quali:

- l'eventualità che il rispondente descrivesse il proprio Titolo di studio anche per le modalità da 01 a 04, sebbene il flusso del questionario non lo richiedesse;
- la possibilità di mancata risposta al quesito precodificato inerente il livello di istruzione e/o il dettaglio sulla durata, nonché l'ulteriore possibilità di incoerenza tra queste risposte ed il testo digitato;
- la probabilità di non esaustività della risposta testuale fornita al fine dell'attribuzione di un codice univoco e al massimo dettaglio.

Per tutti questi motivi, è stato deciso di prevedere dei codici fittizi ovvero codici non contemplati dalla Classificazione ufficiale. La finalità dei codici fittizi è quella di consentire l'individuazione di casistiche particolari da trattare in modo opportuno **nella fase di controllo e correzione;** in pratica, al momento dell'imputazione del codice della classificazione, il codice fittizio permette di selezionarlo tra un sottoinsieme di codici, coerentemente con le informazioni rilevate.

Nel contempo, sono stati anche inseriti nel dizionario *record* per i quali non è stata valorizzata la colonna A. Si è ritenuto, infatti, opportuno contemplare la possibilità di attribuire un codice seppure generico nei casi in cui fosse stata fornita una descrizione non esaustiva del Titolo di studio posseduto e si fosse omesso di rispondere al quesito precodificato.

È il caso per esempio di descrizioni tipo "*qualifica*" a cui può essere associato il codice generico 30000 delle scuole superiori, della durata di 2-3 anni (che non permettono l'accesso all'università),

oppure "*diploma secondo ciclo*" che viene codificato con il codice generico 40000 in quanto riconducibile ad un diploma di istruzione secondaria superiore, della durata di 4-5 anni (che permette l'accesso all'università).

Molti di questi casi sono emersi nella fase di addestramento del sistema e dall'analisi delle risposte fornite nell'Indagine pilota del Censimento 2011.

Si riportano in dettaglio i codici fittizi utilizzati:

- i codici **01999** e **02999** equivalgono a nessun Titolo di studio, rispettivamente "*non sa leggere e scrivere*" e "*sa leggere e scrivere*";
- il codice **93040** indica che, qualora si decida di imputare un Titolo di studio, la scelta dovrà avvenire tra i Titoli di studio con codici che iniziano per 30 (equivalenti ad un corso di studi della durata di 2-3 anni) o tra quelli con codici che iniziano per 40 (equivalenti ad un corso di studi della durata di 4-5 anni);
- il codice **99000** individua studi di natura medico/infermieristica che, dal momento che nel corso degli anni sono stati rilasciati in corrispondenza di diversi livelli di istruzione, costituiscono una casistica particolare la cui soluzione richiede una riflessione ad hoc.

Per comodità di lettura si elencano i Titoli di studio che hanno richiesto, per i motivi esplicitati, l'inserimento dei codici fittizi:

1. codici fittizi relativi alle modalità 01 e 02 del Q. 5.3 (Tavola 9). L'inserimento è stato necessario in quanto dall'analisi dei testi dell'Indagine pilota è emerso che molti rispondenti, sebbene non fosse richiesto, hanno specificato di non possedere alcun Titolo di studio nel Q 5.5;

Tavola 9 - Codici fittizi relativi alle modalità 01 e 02

FILTRO	Codice	F	Titolo di studio
010	01999		NON SA LEGGERE E NON SA SCRIVERE
020	02999		SA LEGGERE E SCRIVERE
020	02999		1 ELEMENTARE
020	02999		2 ELEMENTARE
020	02999		3 ELEMENTARE
020	02999		4 ELEMENTARE

2. codice fittizio 93040, assegnato a tutti quei Titoli di studio con uguale denominazione che, in assenza della durata del corso di studi, possono provenire sia da un corso di studi della durata di 2-3 anni (qualifiche) che da un diploma di stato della durata di 4-5 anni (Tavola 10);

Tavola 10 - Denominazioni assegnate al codice fittizio 93040

FILTRO	Codice	F	Titolo di studio
	93040		Architettura e arredo
	93040		Arte dei metalli
	93040		Arte dei metalli ed oreficeria
	93040		Arte dei rivestimenti ceramici edilizi
	93040		Arte del corallo
	93040		Arte del gres
	93040		Arte del legno
	93040		Arte del legno e restauro del mobile antico
	93040		Arte del merletto e ricamo
	93040		Arte del mobile
	93040		Arte del mosaico
	93040		Arte del tessuto
	93040		Arte del vetro
	93040		Arte del vetro e cristallo
	93040		Arte della ceramica
	93040		Arte della moda e del costume

Tavola 10 segue **- Denominazioni assegnate al codice fittizio 93040**

FILTRO	Codice	F	Titolo di studio
	93040		Arte della porcellana
	93040		Arte della stampa
	93040		Arte dell'alabastro
	93040		Arte dell'arredamento
	93040		Arte delle pietre dure
	93040		Arte grafica pubblicitaria e fotografia
	93040		Arte pubblicitaria
	93040		Arti grafiche
	93040		Calcografia
	93040		Decorazione pittorica
	93040		Decorazione plastica
	93040		Design per ambiente
	93040		Disegnatori architettura e arredamento
	93040		Disegnatori d'architettura
	93040		Disegno animato
	93040		Elettrotecnica e automazione
	93040		Litografia
	93040		Matematico-scientifico
	93040		Oreficeria
	93040		Pittura e decorazione pittorica
	93040		Restauro ceramico
	93040		Rilegatura artistica e restauro del libro
	93040		Rilievo e catalogazione
	93040		Scenotecnica
	93040		Scultura e decorazione plastica
	93040		Tecnico dei servizi sociali
	93040		Xilografia
	93040		DIPLOMA IN ARTE GRAFICA
	93040		LITOGRAFO
	93040		Comunicazioni visive
	93040		Tecnologia ceramica

3. codice fittizio **99000** previsto per tutti gli studi di natura medico/infermieristica che non specificano il livello di studio, oppure è presente una biffatura nel Q.5.3 non compatibile con il Titolo di studio dichiarato (Tavola 11). Come detto sopra, tali Titoli di studio, nel corso degli anni, sono stati rilasciati in corrispondenza di diversi livelli di istruzione.

Tavola 11 - Denominazioni assegnate al codice fittizio 99000

FILTRO	Codice	F	Titolo di studio
	99000		Infermiere
	99000		Diploma di infermiere
	99000		Corso professionale di infermiere
	99000		DIPLOMA DI INFERMIERA
	99000		DIPLOMA DI INFERMIERA PROFESSIONALE
	99000		DIPLOMA DI OSA
	99000		FISIOTERAPISTA
	99000		INFERMIERA PROFESSIONALE
	99000		INFERMIERE
	99000		OPERATORE SOCIO SANITARIO
	99000		OSTETRICIA
	99000		SCUOLA TECNICA DI INFERMIERA
	99000		TECNICO DI RADIOLOGIA MEDICA
	99000		TECNICO OSTETRICO

4.2.5 La fase di addestramento del sistema

Come è stato già ampiamente detto, il dizionario informatizzato per la codifica *batch* è stato arricchito anche con i diversi modi di descrivere i Titoli di studio. Questo lavoro ha richiesto, comunque, un'attenta riflessione ed analisi che ha permesso di individuare i sinonimi più vicini al modo di esprimersi dei rispondenti assicurandosi contemporaneamente che non fossero ambigui.

Anche in questa fase di addestramento del sistema ci si è avvalsi delle potenzialità di ACTR per evidenziare eventuali incoerenze e contemporaneamente ampliare il dizionario con sinonimi ed empiriche. È stato creato a tale scopo un contesto di codifica automatica che utilizza il dizionario della Classificazione ufficiale del Titolo di studio con un campo filtro compatibile con il quesito precodificato previsto nel modello cartaceo del Censimento della Popolazione 2011.

Poiché una parte del lavoro era già stata realizzata nel precedente censimento, si è pensato di sottomettere il dizionario del CP01 (Censimento Popolazione 2001) costituito da 3.202 *record* totali ad un passaggio di codifica *batch* con il contesto preparato con la attuale Classificazione dei corsi di studio. Prima di sottomettere il vecchio dizionario alla codifica è stato però necessario convertire i precedenti campi filtro negli attuali, tenendo conto per esempio che nel precedente censimento i Titoli di studio universitari erano raggruppati soltanto in due modalità di risposta, in quanto non erano contemplate tutte le Lauree di primo e secondo livello del nuovo ordinamento.

Dall'analisi degli *output* di questo test di codifica e soprattutto da quella sui *match* falliti è emersa allora la necessità di inserire nel dizionario testi generici del tipo: "*Diploma di scuola diretta a fini speciali*", "*Diploma parauniversitario*" e sinonimi del tipo: "*Diploma Isef "*, "*Qualifica artigianale*".

Dopo questa prima fase di addestramento e ampliamento del dizionario, è stato effettuato un ulteriore test utilizzando come *file* di *input* le risposte testuali provenienti dall'Indagine pilota sul Censimento della Popolazione 2011. Il *file*, costituito da 8.390 *record*, ha permesso di effettuare diversi passaggi di codifica automatica per evidenziare al meglio le eventuali casistiche da risolvere:

- Passaggio di codifica automatica solo sul campo testo senza l'utilizzo del campo filtro;
- Passaggio di codifica automatica con l'utilizzo del campo filtro;
- Passaggio di codifica automatica solo sulle modalità 06, 07 e 08 senza durata corso di studi.

Anche in questo caso l'analisi successiva degli *output* è risultata determinante per aiutare a risolvere alcuni casi. Dall'esame dei casi di successo nell'attribuzione dei codici relativi a qualifiche, diplomi e lauree triennali è stato confermato il livello di accuratezza atteso. Per quanto attiene invece i Titoli di studio corrispondenti alla modalità 17 (Laurea del vecchio ordinamento, Laurea specialistica o magistrale a ciclo unico del nuovo ordinamento, Laurea biennale specialistica di II livello del nuovo ordinamento), è stato necessario verificare se alcune Lauree specialistiche/magistrali fossero presenti anche nel vecchio ordinamento e, in caso affermativo, inserirne le dizioni. Dai match falliti è emersa infine la necessità di inserire alcune definizioni empiriche, quali, per esempio, "*Capitano lungo corso*" e "*Capitano superiore lungo corso*" all'interno della modalità 09 come "*Diploma di Istituto tecnico ad indirizzo nautico*", in quanto tali descrizioni del Titolo di studio è risultata molto usata dai rispondenti.

4.3 Stato estero

4.3.1 La classificazione degli Stati esteri

La classificazione degli Stati esteri presenta problematiche tutt'altro che banali, non tanto per la complessità della classificazione in sé che, a livello di struttura, è relativamente semplice (le voci non sono moltissime e anche un'eventuale aggregazione per livelli gerarchici prevede un ridotto numero di livelli), quanto piuttosto per la delicatezza delle questioni, soprattutto a livello politico e diplomatico, che sono sottese. Per farsi un'idea si pensi anche soltanto alla spinosissima, estremamente complessa e delicata vicenda dei Territori dell'autonomia palestinese, che è questione ad oggi ancora aperta. Oltre alle problematiche di questa natura, esiste la difficoltà di individuare una classificazione universalmente valida, ossia valida indipendentemente dalla variabile le cui manifestazioni si intende

classificare. Le voci da contemplare o non contemplare nella classificazione, e quindi la classificazione stessa, sono infatti legate a doppia mandata con la finalità per la quale si intende utilizzarla, ovvero con la natura della variabile alla quale si intende applicarla. Questa questione è saldata strettamente con la questione della "storicizzazione" della classificazione. Si pensi ad un utilizzo della classificazione ai fini della rappresentazione delle modalità della variabile cittadinanza, rispetto ad un utilizzo ai fini della rappresentazione delle modalità della variabile luogo di nascita. Poiché in molti casi il quesito sulla variabile luogo di nascita è un quesito retrospettivo, mentre altrettanto non si può dire normalmente per la variabile cittadinanza, è evidente che la distanza tra i due istanti temporali cui si fa riferimento per le due variabili, pur rilevate nello stesso momento, determina di per sé la necessità di disporre di classificazioni diverse, ossia con voci differenti. Uno stato o un territorio possono infatti esistere o essere esistiti in un certo periodo storico, ma non in un altro.

Vi possono poi essere esigenze diverse di classificazione del territorio oltre i confini del proprio stato, in cui la diversità tra le classificazioni idonee non è legata al fattore tempo. Per esempio, mentre per la codifica delle modalità della variabile Cittadinanza della popolazione straniera residente in Italia si ricorre alla Classificazione degli Stati esteri, per classificare per luogo di provenienza il flusso di una merce, piuttosto che un flusso turistico, può essere interessante disporre di una classificazione più ampia, che tenga conto magari di quei territori che possono rivestire per quel particolare tipo di fenomeno un'importanza particolare, legata per esempio alla particolare collocazione geografica, che rende necessarie distinzioni al di sotto del livello di stato.

L'Istat pubblica regolarmente la Classificazione degli Stati esteri esistenti al 31 dicembre di ciascun anno di calendario. Essa contempla gli stati, ovvero quelle entità geografiche e politiche dotate di indipendenza e sovranità, riconosciute dall'Italia e/o a livello internazionale (fanno eccezione i soli Territori dell'autonomia palestinese, che sono anch'essi compresi). La classificazione, nella sua configurazione attuale, risale all'anno 1993. Sul sito dell'Istat è pubblicata per gli anni dal 2002 al 2011 (http://www.istat.it/it/archivio/6747). Esistono anche versioni precedenti il 1993, per le quali tuttavia la struttura, la composizione e i codici erano molto differenti dall'attuale.

Recentemente nel predisporre la classificazione, che nasce come si è detto in un'ottica trasversale (ossia per la rappresentazione di volta in volta del dato dell'anno), si è prestata particolare attenzione all'eventualità di un suo possibile utilizzo di tipo longitudinale, o retrospettivo. Si sono ad esempio adottati criteri più restrittivi per la gestione dei codici, in particolare per quanto riguarda l'attribuzione di nuovi codici agli stati che nascono e la cessazione dei codici assieme agli stati che cessano di esistere. Al momento tuttavia non esiste ancora una classificazione degli Stati esteri "storica", che vada indietro nel tempo oltre quanto indicato sopra. Ad oggi esiste soltanto un esperimento di costruzione di una classificazione storica degli Stati esteri esistiti dal 1918 ad oggi, che però non può essere utilizzata per interpretare i dati degli anni passati, ma solo per una eventuale codifica a ritroso delle informazioni che rispetti un principio di coerenza anche in senso longitudinale.

In occasione del 15° Censimento generale della popolazione e delle abitazioni si è reso necessario approntare una Classificazione degli Stati e dei Territori che rispondesse nei contenuti al dettato del Regolamento (CE) n. 1201/2009 della Commissione, recante attuazione del Regolamento (CE) n. 763/2008 del Parlamento europeo e del Consiglio e riguardante le specifiche tecniche delle variabili e delle relative classificazioni da adottare nella nuova tornata di Censimenti della popolazione e delle abitazioni. Per la prima volta nel 2011, il Censimento è soggetto alla normativa europea, vincolante per gli stati membri in ordine ai contenuti (informazioni da rilevare, definizioni dei relativi concetti e classificazioni), al piano di diffusione (tabelle da produrre e tempi) e alla qualità dei dati prodotti. Lo scopo è assicurare la comparabilità dei risultati dei censimenti della popolazione e delle abitazioni effettuati negli stati membri e garantire l'attendibilità dei documenti di sintesi da compilare a livello comunitario.

La classificazione adottata per il Censimento della popolazione e delle abitazioni del 2011, sulla base delle indicazioni contenute nella normativa europea, rispetto alla classificazione Istat degli Stati esteri, contiene in aggiunta le voci relative ad alcuni territori. Tra le variabili rilevate con il censimento, la classificazione è utilizzata per quelle relative allo stato estero di cittadinanza e di cittadinanza precedente, allo stato o al territorio estero di nascita e a quello di nascita dei genitori, allo stato di ultima residenza, di dimora abituale nel 2010, di dimora abituale nel 2006, di studio o di

lavoro degli individui. Il Regolamento sopra citato fa esplicito riferimento alle due variabili cittadinanza e stato/luogo di nascita.

La classificazione degli Stati esteri e dei Territori per la codifica delle variabili del 15° Censimento della popolazione e delle abitazioni predisposta in conformità con il dettato del Regolamento si compone di 227 voci, a fronte delle 197 previste dalla classificazione Istat degli Stati esteri (compresa la voce "Apolide"). Nel dettaglio, rispetto a quest'ultima, con riferimento alla variabile cittadinanza, la classificazione adottata per il censimento comprende anche i territori del Baliato di Guernsey, di quello di Jersey, dell'Isola di Man, di Sark, di Mayotte, dell'Isola di Sant'Elena, di Anguilla, delle Isole Antille Olandesi, di Aruba, di Bermuda, delle Isole Cayman, delle Isole Falkland (Malvine), dell'Isola di Montserrat, di Saint Martin, di Saint Pierre e Miquelon, di Saint Barthélemy, dei Territori australi e antartici francesi, delle Isole Turks e Caicos, delle Isole Vergini Britanniche, della Nuova Caledonia, dell'Isola di Pitcairn, della Polinesia francese, delle Isole Wallis e Futuna. Si tratta principalmente di dipendenze del Regno Unito, della Francia o dei Paesi Bassi, con maggiore o minore autonomia a seconda dei casi. Viceversa non contempla lo stato del Kosovo, riconosciuto solo parzialmente a livello internazionale, ma riconosciuto dal nostro Paese; non contempla neppure lo stato del Sud Sudan, recentemente costituito, e i Territori dell'autonomia palestinese, al momento non riconosciuti come stato.

In relazione alla variabile Stato/luogo di nascita, nella classificazione viene recuperato il Kosovo e vengono introdotti in aggiunta i territori delle Isole Faer Oer, di Gibilterra, della Groenlandia e i tre stati cessati di Cecoslovacchia, U.R.S.S. e Jugoslavia, la cui codifica a parte nella normativa internazionale è indicata come facoltativa. Non è compresa invece la voce Apolide, presente nella classificazione per la variabile cittadinanza. La classificazione comprende inoltre le voci "Altro" e "Non indicato".

4.3.2 L'aggiornamento dei dizionari degli Stati esteri

Per l'approntamento del dizionario degli Stati esteri e dei Territori per la codifica delle variabili del 15° Censimento della popolazione e delle abitazioni si è partiti dal dizionario utilizzato per il censimento precedente. I territori citati sopra, nella classificazione adottata per il Censimento del 2001, non risultavano enucleati dagli stati di appartenenza e quindi neppure codificati a parte. Nel vecchio dizionario erano elencati semplicemente come sinonimi dei rispettivi stati-padre.

Le operazioni di aggiornamento dei dizionari per la codifica della variabile Stato/Territorio estero, ai fini dell'adeguamento alla nuova normativa internazionale, hanno comportato:

- la revisione completa dei dizionari utilizzati per il Censimento del 2001;
- l'aggiornamento dell'elenco degli stati: aggiunta di stati sorti ed eliminazione di stati cessati nel periodo tra il 21 ottobre 2001 e il 9 ottobre 2011;
- l'enucleazione dei territori da classificare come entità a sé stanti secondo la nuova normativa internazionale;
- l'adeguamento delle descrizioni degli stati e dei territori allo standard delle denominazioni della classificazione degli Stati esteri dell'Istat;
- la traduzione dei dizionari in cinque diverse lingue (limitatamente ad alcune voci, cfr. par. 4.3.4);
- il completamento del dizionario con le voci relative a tutti i sinonimi utili per la corretta codifica dell'informazione a posteriori.

Complessivamente sul dizionario utilizzato per il Censimento del 2001 sono stati effettuati circa 250 interventi, tra inserimenti di nuovi *record* e modifiche di *record* già esistenti. Alla fine il nuovo dizionario è risultato composto di 2.135 voci, rispetto alle 1.932 di quello utilizzato per il censimento precedente.

Oltre al dizionario in lingua italiana, sono stati messi a punto altri quattro dizionari in altrettante lingue estere. I dizionari sono contenuti all'interno di cinque *file* corrispondenti ad altrettanti elenchi di stati e territori esteri espressi in:

- italiano;
- inglese;
- francese;

- tedesco;
- spagnolo.

La struttura dei dizionari è la seguente, laddove alcune informazioni (quelle riportate nei campi/colonne C e D) non sono utilizzate per il *matching* testuale, ma, come si vedrà di seguito, per la visualizzazione nel caso di compilazione del questionario elettronico.

- **Colonna A**: sono riportati i codici della classificazione ufficiale Istat;
- **Colonna B**: sono riportate le descrizioni corrispondenti agli stati esteri su cui effettuare la ricerca testuale;
- **Colonna C**: se valorizzata, riporta se la descrizione corrispondente si riferisce ad un *continente*, ad un *territorio*, ad una *città* diversa dalla capitale;
- **Colonna D**: se valorizzata individua la denominazione ufficiale.

Entrando nel dettaglio, nella **Colonna A** sono riportati i codici della classificazione ufficiale Istat, cui sono stati aggiunti:

- il codice Italia = 100;
- i 26 codici relativi ad altrettanti territori da classificare separatamente rispetto agli stati di appartenenza, in base alle raccomandazioni internazionali;
- alcuni codici riferiti a paesi cessati o soppressi (esempio URSS, Cecoslovacchia, Jugoslavia, ecc.);
- il codice relativo al "Non indicato";
- cinque nuovi codici riferiti ai continenti, qualora la risposta fornita consenta soltanto l'individuazione del continente e non dello stato (Europa = 1, Africa = 2, Asia = 3, America = 4, Oceania = 5).

Nella **Colonna B** sono riportate le descrizioni corrispondenti agli Stati o ai Territori esteri. Le denominazioni utilizzate sono quelle ufficiali Istat. La classificazione è stata inoltre integrata con altre denominazioni e forme identificative degli Stati esteri (dette anche sinonimi) che possono consentire di recuperare l'informazione anche quando questa sia fornita in maniera impropria. In particolare sono state inserite:

- alcune denominazioni di uso locale (per es. *Bosnia i Hercevogina* per Bosnia-Erzegovina, *Al Magrib* per Marocco, *Rossijskaja Federacija* per Federazione Russa);
- alcune denominazioni non più in uso (per es. *Birmania* per Myanmar, *Zaire* per Repubblica Democratica del Congo, *Ceylon* per Sri Lanka);
- alcune denominazioni alternative a quella ufficiale Istat (per es. *Russia Bianca* per Bielorussia);
- alcune denominazioni riferite a paesi soppressi in seguito ad unificazione, a cui è stato attribuito il codice dell'attuale paese di appartenenza:
- DDR / RDT / Repubblica Democratica Tedesca = Germania (cod. = 216);
- Yemen del Nord / Yemen del Sud = Yemen (cod. = 354);
- Vietnam del Nord / Vietnam del Sud = Vietnam (cod. = 353).
- alcune denominazioni riferite a territori, possedimenti o parti degli Stati Esteri;
- il nome della capitale e di alcune altre città;
- i codici ISO degli stati;
- gli aggettivi maschile e femminile di cittadinanza.

Nella **Colonna C,** se valorizzata, è riportato se la descrizione corrispondente si riferisce ad un *continente*, ad un *territorio*, ad una *città* diversa dalla capitale**.**

La **Colonna D**, come già detto, se valorizzata indica che la descrizione di colonna B corrisponde alla denominazione ufficiale del paese.

In sintesi, per ciascuno Stato estero o Territorio da classificare separatamente secondo le raccomandazioni internazionali, sono fornite le informazioni di cui alla Tavola 12:

Tavola 12 - Descrizioni associate a ciascuno stato nel dizionario

COLONNA A	COLONNA B
COD Istat	sigla ISO
COD Istat	Ex nome (eventuale)
COD Istat	denominazione ufficiale
COD Istat	nome/i alternativi (eventuale altra denominazione)
COD Istat	aggettivo maschile di nazionalità
COD Istat	aggettivo femminile di nazionalità
COD Istat	aggettivo/i maschile di nazionalità (eventuale altro/i)
COD Istat	aggettivo/i femminile di nazionalità (eventuale altro/i)
COD Istat	nome della capitale
COD Istat	nome della capitale (eventuale altra)
COD Istat	sigla ISO

Come già accennato, nel *file* sono presenti inoltre i *record* che riportano le denominazioni dei continenti, nonché quelli dei territori e di altre città, associati al codice dello stato cui si riferiscono.

Tavola 13 - Ulteriori denominazioni incluse nel dizionario

COLONNA A	COLONNA B
COD Istat	nome del continente
COD Istat	Aggettivo maschile di continente
COD Istat	Aggettivo femminile di continente
COD Istat	nome di possedimento, territorio
COD Istat	nome di possedimento, territorio (eventuale altro/i)
COD Istat	nome di possedimento, territorio (eventuale altra/e lingua/e)
COD Istat	nome di città diversa dalla capitale
COD Istat	nome di città diversa dalla capitale (eventuale altra/e)

4.3.3 Il dizionario per la codifica con il questionario Web

Per la compilazione del questionario *Web* è stato messo in linea esclusivamente il dizionario in italiano, in quanto si è ritenuto che la possibilità di visualizzare questo elenco di stati e territori esteri desse sufficienti garanzie per la corretta codifica.

Inoltre, per il questionario *online,* è stato utilizzato per il *display* anche il contenuto delle colonne C, in cui, quando valorizzato, è riportato se la descrizione corrispondente si riferisce ad un *continente*, ad un *territorio*, ad una *città* diversa dalla capitale, nonché il contenuto della colonna D che, quando valorizzato, individua la denominazione ufficiale. I territori che sono stati enucleati rispetto allo stato di appartenenza, in base alla normativa internazionale, presentano indicazione dello stato di riferimento, per gli altri compare la dicitura "territorio". Questi ultimi sono codificati con il codice dello stato di appartenenza. Il *display* di queste informazioni è importante in quanto, se il *matching* testuale avviene con una denominazione non ufficiale (un sinonimo), il rispondente ha conferma del fatto che la sua risposta viene ricondotta alla corrispondente denominazione ufficiale.

4.3.4 Il dizionario per la codifica in batch di quanto rilevato con il questionario cartaceo

Per la codifica in *batch* dei dati rilevati con il questionario cartaceo, sono stati resi disponibili anche i dizionari nelle lingue diverse dall'italiano.

Relativamente a questi dizionari (in inglese, francese, tedesco e spagnolo), si fa presente che, sebbene la struttura sia analoga all'elenco in lingua italiana (a meno delle colonne C e D), questi contengono una minore varietà di dizioni rispetto a quello in italiano: contengono solo le descrizio-

ni standard e non sono integrati, come invece avviene nel *file* in italiano, né da descrizioni diverse da quelle ufficiali e/o correnti né da nomi di città diverse dalla capitale. Non sono inoltre presenti i codici relativi ai continenti, ai territori e/o dipendenze che non siano da classificare separatamente rispetto agli stati di appartenenza, in base alle raccomandazioni internazionali. La loro struttura è quindi quella riportata nella seguente tavola.

Tavola 14 - Descrizioni associate a ciascuno stato nei dizionari utilizzati per la codifica in *batch*

COLONNA A	COLONNA B
COD Istat	sigla ISO
COD Istat	ex nome
COD Istat	denominazione dello stato (una riga ulteriore in caso di altra denominazione lunga - eccetto per i territori e gli ex-nomi)
COD Istat	aggettivo di nazionalità (una riga ulteriore in caso di altro aggettivo di nazionalità)
COD Istat	nome della capitale (una riga ulteriore in caso di altra città)

5. Le soluzioni per il questionario elettronico

Una delle principali innovazioni del 15° Censimento della Popolazione è stata la possibilità per tutti i cittadini di poter compilare il questionario *online*. L'Istat ha realizzato una applicazione *Web* che riproduce i due questionari del censimento per le famiglie (versione *long* e *short*) e quello delle convivenze. L'accesso da parte del cittadino avviene utilizzando il proprio codice fiscale e una *password* distribuita insieme al questionario cartaceo. Sono stati ricevuti quasi 8.500.000 questionari via *Web*, pari a oltre il 30% dei questionari totali.

Il principale vantaggio della modalità *online* di compilazione del questionario rispetto a quella cartacea risiede ovviamente nella possibilità di effettuare dei controlli sulla qualità dei dati inseriti prima che questi vengano inviati all'Istat. In questo senso va considerato anche l'utilizzo di dizionari statistici che, nell'ambito di domande che nel questionario cartaceo prevedono l'inserimento di testo libero, consentono invece la scelta di risposte da elenchi. Se da un lato i vantaggi di questa modalità sembrano evidenti, va però sottolineato che si possono ottenere solo se i dizionari sono integrati nell'ambito di un progetto coerente che tenga conto di problematiche di usabilità lato utente e di performance generale.

In questa sezione viene descritto l'utilizzo dei dizionari statistici nell'ambito del questionario *online*, presentando le diverse tipologie di domande e gli aspetti tecnici alla base del progetto.

5.1 Tipologie di domande basate su dizionario

La struttura del progetto del questionario della popolazione è basata su metadati che descrivono le domande e il flusso di compilazione, ovvero la sequenza in cui le domande vanno compilate in base alle risposte date. Queste informazioni sono memorizzate su una base di dati relazionale e caricate in memoria al momento dell'esecuzione dell'applicazione *Web*.

Le domande sono organizzate in diverse tipologie, che definiscono il tipo di controlli di *input* utilizzato (ad esempio, *radio button* o *check box* a risposta multipla, etc.), la formattazione e il funzionamento generale. Questo permette di riutilizzare il codice scritto per una tipologia di domanda in diversi punti del questionario.

Nel seguito di questa sezione si analizzerà il funzionamento delle tipologie di domande del questionario del censimento della popolazione che utilizzano dizionari, ovvero:

- Inserimento di Comune e Provincia;
- Inserimento dello Stato estero;
- Inserimento della denominazione esatta del Titolo di studio.

Come detto, ogni tipologia corrisponde a diverse domande effettive (che a loro volta compaiono nel questionario sia in versione *long* che *short*), elencate nella seguente tavola (il numero della domanda fa riferimento alla versione *long* del questionario):

Tavola 15 - Tipologie di domande e domande del questionario

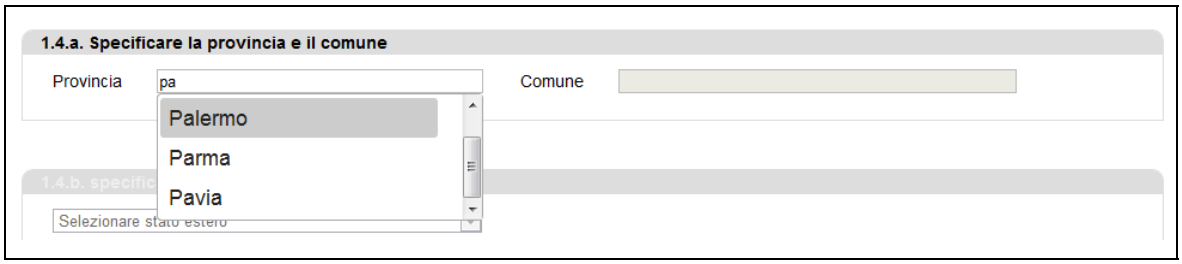
TIPOLOGIA DOMANDA	Domanda
Provincia-Comune	1.4a - luogo di nascita (altro Comune italiano) 4.5a - dimora abituale un anno fa (altro Comune italiano) 4.6a - dimora abituale cinque anni fa (altro Comune italiano) 7.2a - luogo abituale di studio o di lavoro (altro Comune italiano)
Stato estero	3.1a - cittadinanza (Stato estero) 3.3a - Stato estero di cittadinanza precedente 3.4a - Stato estero di nascita padre 3.5a - Stato estero di nascita madre 4.4 - Stato estero ultima residenza 4.6b - dimora abituale cinque anni fa (Stato estero) 7.6b - luogo abituale di studio o di lavoro (Stato estero)
Titolo di studio	5.5 - denominazione esatta Titolo di studio

Per ogni tipologia di domanda verrà presentato il funzionamento lato utente, discutendo le scelte alla base del progetto dell'interfaccia, per poi soffermarsi in particolare sul progetto della tipologia Titolo di studio, che presenta gli aspetti più originali e complessi.

5.1.1 Comune-provincia

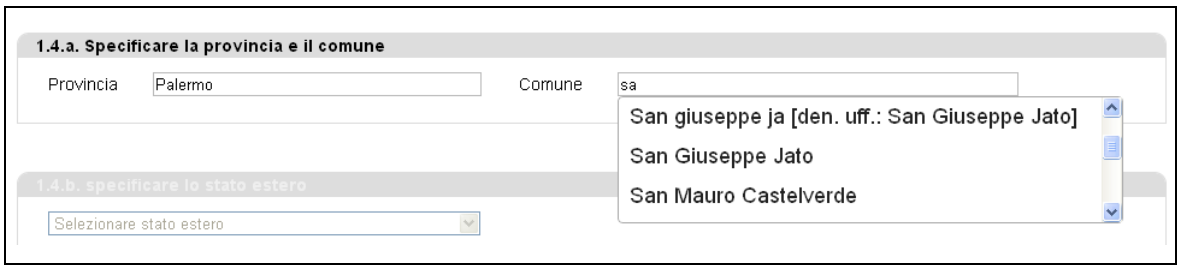
La domanda si presenta come due caselle di testo separate in cui devono essere indicati rispettivamente la Provincia e il Comune. Quando l'utente inizia a scrivere la denominazione della Provincia viene visualizzato un elenco con tutte le Province corrispondenti al testo inserito (Figura 3).

Figura 3 - Scelta della Provincia



Per rispondere effettivamente alla domanda l'utente deve selezionare il nome esatto tra quelli proposti dall'elenco. Una volta scelta la provincia, viene attivata la casella del Comune, dove l'utente ripete lo stesso procedimento (Figura 4). I Comuni visualizzati nel secondo elenco sono ovviamente solo quelli appartenenti alla Provincia precedentemente scelta.

Figura 4 - Scelta del Comune



Il dizionario di Comuni e Province è stato organizzato in modo da rispondere al problema di consentire all'utente di identificare correttamente il Comune e la Provincia anche nei frequenti casi di cambio di denominazione e/o di variazione territoriale avvenuti nel tempo. Il problema è reso più complesso dal fatto che le varie domande in cui si richiede il Comune si riferiscono a tempi differenti: ad esempio il luogo abituale di studio e di lavoro deve far riferimento alla situazione al

9 ottobre 2011, mentre per il luogo di nascita devono essere considerate tutte le possibili situazioni dal 1895 a oggi.

Per gestire tutti i diversi casi, il dizionario include tutti i Comuni, con relative variazioni di denominazione e territorio, a partire dal 1895. Sono quindi compresi anche Province (e relativi Comuni) non più appartenenti al territorio italiano (Fiume, Pola e Zara). Inoltre, sono presenti anche i nomi di frazioni e località nonché diverse scritture per uno stesso Comune. Per ogni denominazione, nel dizionario viene indicato se è quella corrente del Comune in questione o se è relativa ad una situazione precedente. In questo modo, nelle domande che fanno riferimento alla data del censimento è possibile determinare e mostrare all'utente solo le denominazioni attive.

Complessivamente, il dizionario contiene circa 16.000 denominazioni di Comuni. Le denominazioni di Comuni e Province non sono tradotte nel passaggio alle altre lingue supportate dall'applicazione (tedesco e sloveno), anche se per alcuni Comuni è presente la doppia denominazione (in particolare per i Comuni della provincia di Bolzano/Bozen).

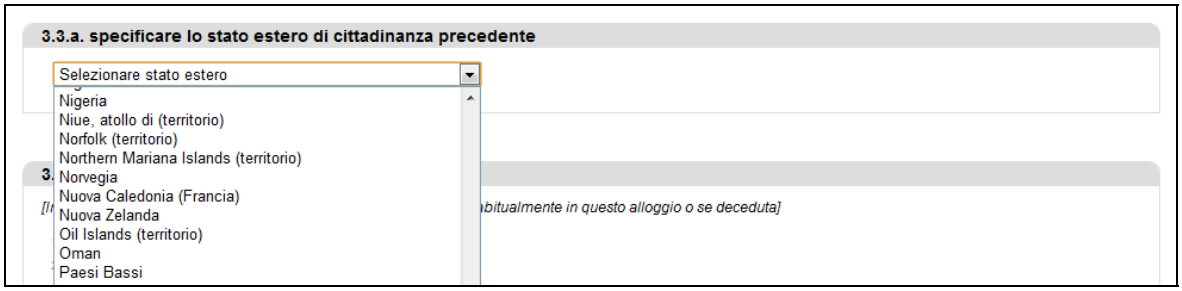
La scelta di un inserimento separato tra Provincia e Comune è stata dettata proprio da considerazioni legate alla dimensione del dizionario. Infatti, anche se per l'utente poteva essere più agevole inserire direttamente il Comune (e ottenere la Provincia di conseguenza), questa modalità avrebbe comportato la visualizzazione di elenchi troppo lunghi. Attraverso questa scelta, si è anche ottimizzata la performance e la scalabilità perché si sfrutta un meccanismo di indicizzazione dei Comuni tramite il codice della Provincia, che dà luogo a una ricerca negli elenchi estremamente efficace.

5.1.2 Stato estero

Il dizionario degli Stati esteri contiene sia le denominazioni ufficiali degli Stati esteri attuali che denominazioni di stati non più esistenti e di altri territori che non hanno caratteristica di stato autonomo (ad esempio "Scozia").

A livello di interfaccia utente si è scelto di presentare le possibili risposte organizzate in una lista a discesa (Figura 5) piuttosto che prevedere un autocompletamento come nel caso precedentemente analizzato. Il motivo è legato a possibili discordanze tra le denominazioni ufficiali degli stati e i nomi comunemente utilizzati (anche dovute ad esempio a problemi di traduzione) che potevano rendere più difficile identificare correttamente uno stato.

Figura 5 - Scelta dello Stato estero



La versione completa del dizionario contiene 412 denominazioni. Per la codifica *online*, sono stati predisposti anche un dizionario in lingua tedesca e uno in lingua slovena.

5.1.3 Titolo di studio

La domanda 5.5 prevede l'inserimento da parte dell'utente della denominazione estesa del Titolo di studio conseguito. Laddove questa domanda nel questionario cartaceo prevede l'inserimento di testo libero, nel questionario *online* si è scelto di richiedere la selezione della risposta da un dizionario.

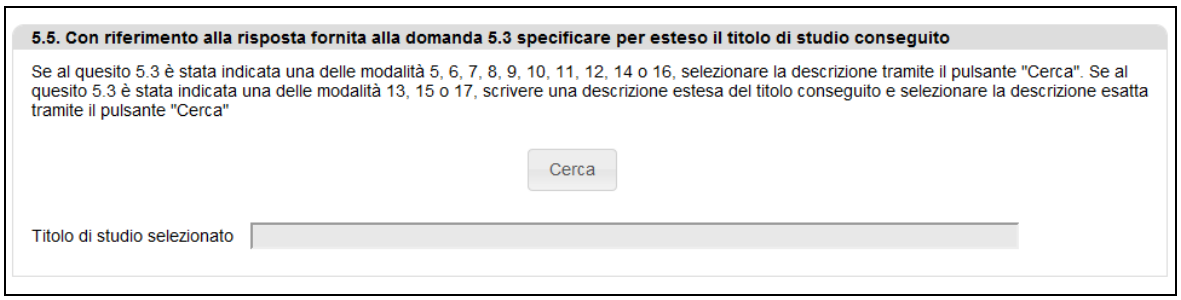
Rispetto alle altre tipologie di domande, però, ha delle specificità da considerare nel progetto. Il dizionario completo contiene un numero molto alto (6625) di elementi le cui denominazioni sono spesso difficili da distinguere tra loro e da inquadrare rispetto alle frequenti variazioni legislative e organizzative, soprattutto a livello universitario. Inoltre, il Titolo di studio indicato deve essere coe-

rente con le risposte date alle domande 5.3 e 5.4, nelle quali si deve indicare il Titolo di studio più elevato conseguito, scegliendo tra un elenco e, talvolta, la durata del corso di studi. I casi più complessi da codificare sono quelli afferenti alle modalità corrispondenti a titoli universitari (13, 15, 17), che sono collegate ad una serie di denominazioni la cui numerosità (rispettivamente 272, 1.773 e 4.511 elementi) è tale da non rendere pratico un elenco fisso da cui scegliere. Per i Titoli di studio che fanno riferimento alle altre modalità sono stati, invece, predisposti dei semplici menù a tendina.

L'insieme di tali requisiti ha richiesto la realizzazione di una soluzione articolata che non è riconducibile a nessun componente di interfaccia *standard* e che incorpora un vero e proprio motore di ricerca dei Titoli di studio. Il progetto della domanda verrà analizzato in dettaglio nel seguito, presentando in questa sezione il funzionamento dell'interfaccia utente e nella successiva soffermandosi sugli aspetti progettuali relativi al motore di ricerca.

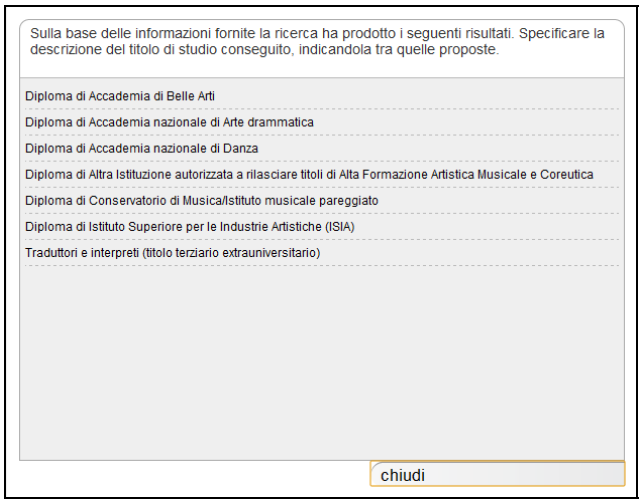
L'interfaccia della domanda 5.5 varia in base alla risposta data alla domanda 5.3. In primo luogo la domanda si attiva solo se si risponde di aver conseguito un Titolo di studio superiore (modalità da 5 in poi). Se la risposta alla domanda 5.3 è compresa tra le modalità 5 e 16, escludendo 13 e 15, la domanda 5.5 si presenta inizialmente all'utente con attivo un pulsante per la ricerca del Titolo di studio esatto (Figura 6).

Figura 6 - Scelta del Titolo di studio



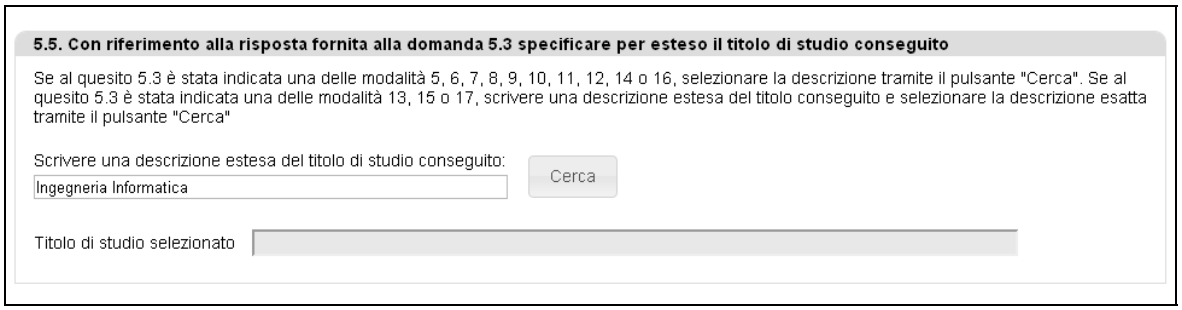
Premendo il pulsante "Cerca" viene visualizzato un elenco di titoli di studio compatibili con la risposta 5.3, tra cui l'utente deve effettuare la sua selezione. Ipotizzando che l'utente abbia selezionato la modalità 12 ("*Diploma di Accademia di Belle Arti, Danza, Arte Drammatica, ISIA, ecc. Conservatorio (vecchio ordinamento)*") alla domanda 5.3, verrebbe visualizzato l'elenco mostrato in figura 7.

Figura 7 - Scelta del Titolo di studio



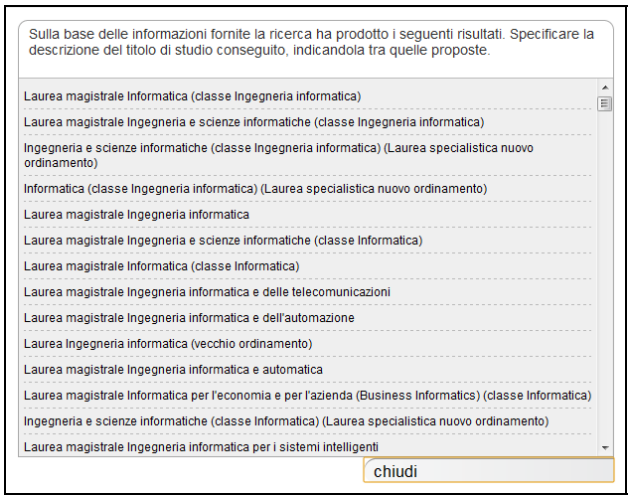
Nel caso in cui invece la risposta alla domanda 5.3 sia una tra le modalità 13, 15 o 17, come detto, l'elenco delle denominazioni collegate ai Titoli di studio assume dimensioni troppo grandi e fornire un elenco prestabilito non è una soluzione fattibile né dal punto di vista dell'usabilità né da quello delle prestazioni. Perciò in questi casi viene mostrato accanto al tasto "Cerca" un campo di ricerca in cui l'utente può inserire una descrizione estesa del Titolo di studio che viene usata come chiave dal motore di ricerca (Figura 8).

Figura 8 - Scelta del Titolo di studio



Premendo "Cerca", il testo digitato viene quindi sottoposto al motore di ricerca, che genera un elenco delle denominazioni ufficiali che maggiormente corrispondono al testo cercato, applicando un meccanismo che verrà descritto nel dettaglio nelle prossime sezioni. Ipotizzando che l'utente alla domanda 5.3 abbia selezionato la modalità 17 ("*Laurea (4-6 anni) del vecchio ordinamento, laurea specialistica o magistrale a ciclo unico del nuovo ordinamento, laurea biennale specialistica (di II livello) del nuovo ordinamento*") e alla domanda 5.5 abbia digitato il testo "Ingegneria Informatica" verrebbe visualizzato l'elenco mostrato in figura 9.

Figura 9 - Scelta del Titolo di studio



5.2 La fase di pre-elaborazione *offline* del dizionario dei titoli di studio

Prima di poter essere utilizzati dal motore di ricerca del questionario *online*, le voci presenti nel dizionario dei Titoli di studio (relative alle modalità 13, 15 o 17 della domanda 5.3) devono essere pre-elaborate da una procedura che consenta al sistema di accedere alle informazioni in modo rapido ed efficace. Essendo relativamente pesante dal punto di vista computazionale, tale procedura viene eseguita, una volta per tutte e in modalità *offline*, a partire dalla versione definitiva del dizionario. La riesecuzione della procedura si rende necessaria solo in caso di modifica delle voci presenti nel dizionario. Schematicamente i passi fondamentali della procedura di pre-elaborazione *offline* sono i seguenti:

a) normalizzazione dei caratteri presenti nelle voci di dizionario;
b) rimozione dalle voci di dizionario delle *stopword* specifiche;
c) estrazione dei singoli termini (parole) di ricerca dalle versioni normalizzate delle voci di dizionario.

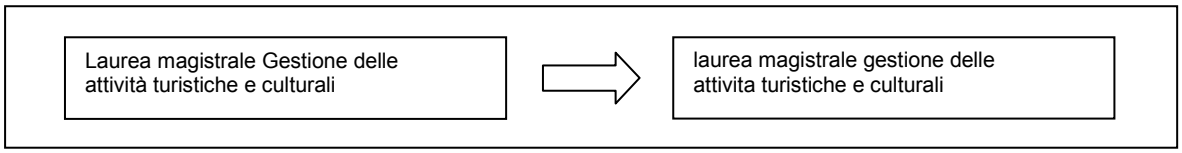
a) Normalizzazione dei caratteri presenti

Nella fase di normalizzazione dei caratteri, a partire da ogni singola voce presente nel dizionario ne viene costruita una nuova versione (normalizzata) in cui:

- tutti i caratteri maiuscoli sono rimpiazzati dal corrispondente carattere minuscolo;
- tutte le vocali accentate sono rimpiazzate con le corrispondenti vocali prive di accento;
- segni di punteggiatura e caratteri speciali come parentesi, trattini, barre, etc. sono rimpiazzati dal carattere spazio;
- (nel caso di lingue straniere) altri caratteri speciali come le consonanti con diacritico vengono rimpiazzate dalla corrispondente consonante priva di diacritico (ad esempio la lettera "š" è rimpiazzata dalla lettera "s" oppure la ß, *scharfes es* tedesca, da una doppia "s").

Al termine della fase eventuali spazi doppi, tripli, etc. derivanti dalla rimozione di alcuni caratteri vengono sostituiti da spazi singoli (cfr. Figura 10)

Figura 10 - Esempio di normalizzazione dei caratteri



b) Rimozione delle stopword

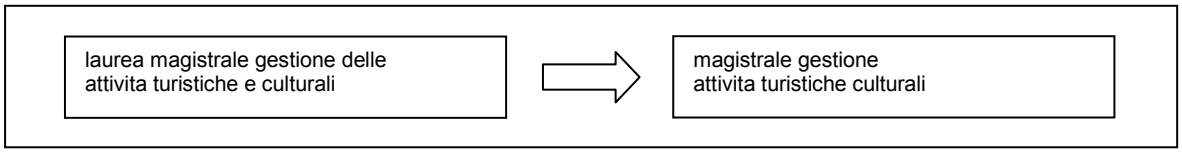
Con il termine inglese *stopword* si indicano parole che possono essere considerate "inutili" limitatamente ai fini dell'elaborazione di un certo frammento testuale. Chiaramente l'"inutilità" di una parola dipende strettamente dal contesto di riferimento e dalle finalità dell'elaborazione. Tipicamente nell'ambito dei motori di ricerca vengono considerate *stopword* le congiunzioni, gli articoli, le preposizioni semplici o articolate, etc. Tuttavia, come detto, l'inutilità di una parola non è un concetto assoluto e dipende spesso dallo specifico contesto considerato. Così in generale il termine "laurea" è tutt'altro che inutile se la ricerca viene svolta su un insieme di documenti/voci di dizionario del tutto generale, ma lo diventa del tutto se la ricerca viene svolta su un dizionario che è costituito esclusivamente da denominazioni di corsi di laurea.

Nell'ambito del motore di ricerca per i Titoli di studio corrispondenti alle modalità 13, 15 o 17 della domanda 5.3 (Titolo di studio più elevato conseguito), sono state preparate delle liste di *stopword* distinte in funzione della modalità scelta dal rispondente. Nella fattispecie sono state preparate le seguenti tre liste:

- una lista di *stopword* generali, costituite essenzialmente da congiunzioni, preposizioni, etc., da applicare su tutti i titoli di studio in esame;
- una lista di *stopword* specifiche per i Titoli di studio relativi alla modalità 13: Diploma universitario (2-3 anni) del vecchio ordinamento (incluse le scuole dirette a fini speciali o parauniversitarie);
- una lista di *stopword* specifiche per i titoli di studio relativi alla modalità 15: Laurea triennale (di I livello) del nuovo ordinamento.

Premesso quanto sopra, la rimozione delle *stopword* consiste nell'eliminare dalla versione normalizzata delle voci di dizionario (ottenuta al passo precedente) tutte quelle parole che coincidono con *stopword* specifiche relativamente a quella voce. Ad esempio, dalle voci riferite alla modalità 15 vengono eliminate sia quelle *stopword* che risultano valide per qualunque modalità (congiunzioni, preposizioni, etc.) sia quelle specifiche della sola modalità 15 (cfr. Figura 11).

Figura 11 - Esempio di rimozione delle *stopword*

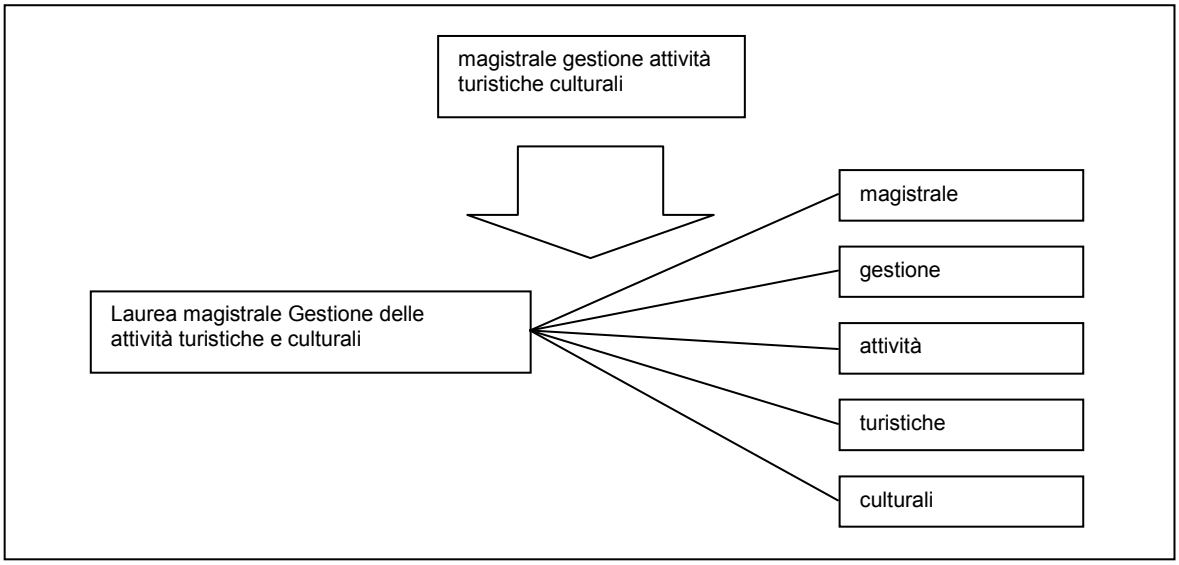


c) Estrazione dei termini di ricerca

Una volta costruita una versione normalizzata delle voci di dizionario si rende necessario estrarre i singoli termini (parole) che lo costituiscono, in quanto la ricerca *online* agirà sulle singole parole e non sulla voce di dizionario nel suo complesso (cfr. Figura 12). Una volta eseguiti i passi di normalizzazione descritti nei passi precedenti, questa estrazione è piuttosto semplice, in quanto richiede semplicemente di estrarre le singole parole componenti la voce, usando il carattere di "spazio" come delimitatore di parola.

Le singole parole (termini di ricerca) così ottenute vengono a questo punto salvate nella base dati del questionario, inserendo il singolo termine (se non è già stato trovato in altra voce) e inoltre un legame (relazione) tra la voce di dizionario e il termine stesso. Se la voce è già presente viene incrementato un valore numerico associato ad ogni termine che rappresenta la frequenza del termine stesso nel dizionario. La frequenza di un termine è importante per il successivo ordinamento dei risultati a fronte di una ricerca da parte dell'utente, in quanto valori alti di frequenza indicano un termine molto comune e quindi una bassa selettività del termine stesso. Viceversa bassi valori di frequenza indicano un termine molto specifico e la sua presenza all'interno di una voce di dizionario dovrebbe essere maggiormente "premiata".

Figura 12 - Esempio di estrazione dei termini di ricerca



A partire dalla sua versione normalizzata e ripulita dalle *stopword*, la voce di dizionario viene collegata ai singoli termini normalizzati, che verranno poi utilizzati dalla procedura del motore di ricerca

5.3 La procedura del motore di ricerca per i Titoli di studio

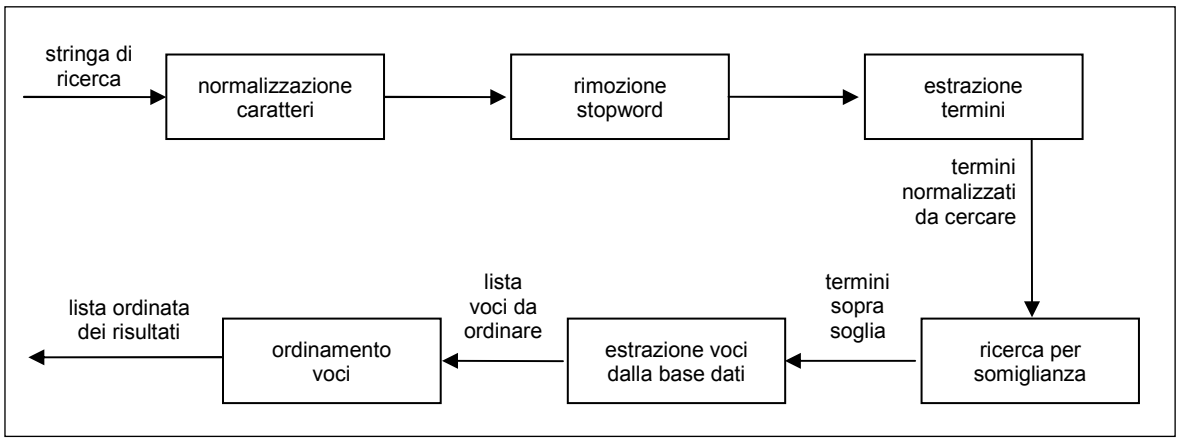
In termini molto generali l'*input* di questa procedura è costituito da una stringa assolutamente generica inserita dall'utente nell'apposita casella di testo e dalla modalità di risposta selezionata dall'utente alla domanda 5.3 del questionario. L'*output* è costituito da una lista di voci di dizionario (nella loro versione "completa" non normalizzata) che corrispondono alla modalità di risposta selezionata e sono sufficientemente "vicine" alla stringa di ricerca specificata dall'utente. L'obiettivo dell'algoritmo alla base del motore di ricerca è fare in modo che tale lista proponga le voci secondo un ordine che privilegi (disponendole ai primi posti dell'elenco) le voci più vicine a quanto espresso

dall'utente, “premiando” in modo particolare le corrispondenze esatte e le corrispondenze con termini poco frequenti nel dizionario (termini rari e quindi altamente selettivi). Di fatto la modalità di risposta alla domanda 5.3 genera un primo filtro sulla lista delle voci candidate, la stringa specificata dall'utente genera un ulteriore filtro e viene inoltre utilizzata dall'algoritmo per stabilire un ordinamento sulla lista da proporre all'utente.

Schematicamente i passi fondamentali che descrivono la procedura utilizzata dal motore di ricerca sono i seguenti (cfr. Figura 13):

a) normalizzazione dei caratteri presenti nella stringa di ricerca;
b) rimozione dalla stringa di ricerca delle *stopword* specifiche;
c) estrazione dei singoli termini (parole) di ricerca dalla versione normalizzata della stringa di ricerca;
d) ricerca per somiglianza dei singoli termini presenti nella stringa di ricerca normalizzata e costruzione della lista di termini sopra soglia di somiglianza;
e) estrazione dalla base dati delle voci di dizionario che contengono uno o più termini presenti nella lista di termini sopra soglia di somiglianza;
f) ordinamento delle voci di dizionario estratti al passo precedente.

Figura 13 - I passi del motore di ricerca per i Titoli di studio



a-c) Normalizzazione, rimozione delle stopword e estrazione dei termini

Le prime tre operazioni sono del tutto analoghe a quelle illustrate in precedenza per la fase di pre-elaborazione *offline*. Diversamente dalla pre-elaborazione, in questo caso le tre operazioni hanno un impatto praticamente trascurabile, in quanto la normalizzazione ed estrazione non coinvolge l'intero dizionario, ma la singola stringa di testo inserita dall'utente. Al termine delle prime tre operazioni si ottiene quindi una lista di termini (parole) in forma normalizzata da utilizzare per i passi seguenti

d) Ricerca dei termini sopra soglia di somiglianza

Per ogni termine presente nella lista dei termini determinata al passo precedente viene calcolata la somiglianza con ciascun termine presente nella base dati del questionario (la tabella dei termini è stata costruita dalla procedura offline descritta in precedenza). Entrano a far parte della lista dei termini sopra soglia tutti e soli quei termini della base dati per i quali la somiglianza con uno dei termini cercati risulta essere superiore ad un valore di soglia prefissato. Tale valore di soglia può essere modificato liberamente e deve essere oggetto di un'attenta attività di calibrazione: valori troppo bassi possono restituire un numero eccessivamente alto di corrispondenze, mentre valori troppo alti possono di fatto restringere i risultati alle sole corrispondenze esatte. La somiglianza è rappresentata da un numero decimale normalizzato a 1, dove il valore unitario si ottiene solo in caso di perfetta uguaglianza e valori via via più bassi indicano una somiglianza progressivamente inferiore.

L'algoritmo di somiglianza adottato si basa su una versione normalizzata della misura di distanza di Jaro-Winkler (Jaro 1989; Winkler 1990), così come implementata nella libreria *open-source* simMetrics, e il valore di soglia utilizzato per il questionario *online* è di 0.95. In altre parole se confrontando uno dei termini (in versione normalizzata) presenti nella stringa di ricerca con uno dei termini nella base dati si ottiene una similarità superiore a 0.95, il termine del dizionario viene incluso nella lista dei termini sopra soglia (ed entrerà quindi nella fase di estrazione del passo successivo).

Si noti che l'uso della misura di similarità (in luogo di una pura e semplice corrispondenza esatta tra termini) ha il duplice vantaggio di fornire una corrispondenza positiva anche in presenza di piccoli errori di battitura (inversioni tra due caratteri consecutivi, omissioni di un carattere, etc.) e di includere anche termini che sono strettamente legati a quello richiesto (ad esempio di includere il termine "psicologica" anche se l'utente ha scritto "psicologia" oppure, erroneamente, "psiclogica"). Di contro il valore di soglia deve essere comunque mantenuto abbastanza alto per evitare l'insorgenza di un elevato numero di "falsi match" ovvero di corrispondenze con termini che non hanno in effetti alcun legame con quanto richiesto dall'utente.

e) Estrazione dalla base dati delle voci di dizionario corrispondenti

Sulla base della lista di termini costruita al passo precedente e delle relazioni tra termini e voci di dizionario determinate nella fase di pre-elaborazione *offline*, vengono estratte dalla base dati tutte le voci di dizionario che hanno uno o più legami con i termini sopra soglia (e quindi in pratica che contengono, nella versione normalizzata, almeno uno di tali termini). Contestualmente a tale estrazione vengono letti o calcolati anche una serie di dati accessori che consentono di ordinare le voci della lista in modo da privilegiare quelle che meglio corrispondono a quanto espresso dall'utente nella propria stringa di ricerca.

In particolare vengono letti/calcolati i dati seguenti:

- per ogni voce di dizionario della lista il numero di termini che costituiscono complessivamente la voce stessa: a parità di numero di corrispondenze vengono preferite le voci che contengono meno termini;
- indice di frequenza dei singoli termini: l'indice si basa sui valori di frequenza determinati durante la pre-elaborazione *offline* e valori prossimi a 1 indicano termini molto rari, mentre valori verso lo zero corrispondono a termini molto comuni;
- numero di corrispondenze esatte (similarità pari a 1) tra termini della stringa di ricerca e termini presenti nella voce: in generale le corrispondenze esatte vengono "premiate" rispetto alle corrispondenze in cui si ha solo un'elevata somiglianza.

f) Ordinamento dei risultati

Sulla base dei dati accessori determinati al passo precedente, a ciascun elemento della lista di voci estratte viene associato un punteggio, sulla base del quale la lista viene ordinata e proposta all'utente attraverso l'interfaccia *Web*. In termini puramente qualitativi i criteri alla base della costruzione di tale punteggio (che si traducono di fatto in coefficienti "premianti" o "penalizzanti") sono i seguenti:

- ogni corrispondenza con un termine deve incrementare il punteggio: quindi una voce che contiene sia il termine T1 che il termine T2 deve ricevere più contributi di una che contiene solo il termine T1 o solo il termine T2;
- se due voci hanno lo stesso numero di corrispondenze con termini di ricerca e in particolare con i medesimi termini è da preferire la voce "più breve" (che contiene complessivamente meno termini);
- a parità di numero di corrispondenze è da preferire la voce che contiene un termine "raro" (bassa frequenza) rispetto a quella che contiene un termine comune (molto frequente);
- se due voci corrispondono al medesimo termine, la voce che ha una corrispondenza esatta (similarità pari a 1) è da preferire a quella che ha soltanto un'elevata somiglianza (similarità strettamente minore di 1).

6. La soluzione per il questionario cartaceo

Come anticipato nel capitolo 2, nel caso di compilazione del questionario cartaceo, la codifica è una delle attività delegate alla società incaricata di implementare le procedure per l'acquisizione con la lettura ottica o con il *data entry*. A tal fine, per garantire un elevato livello di qualità dei codici attribuiti, nonché coerenza rispetto alla codifica effettuata in fase di rilevazione con il questionario elettronico, sono state fornite le basi informative per ciascuna delle classificazioni trattate, così come descritto nel capitolo 3.

D'altra parte, ciascuna delle classificazioni presenta delle peculiarità che impattano direttamente sulla procedura di attribuzione del codice; nel caso dello Stato estero, per esempio, la risposta potrebbe essere fornita in diverse lingue, per cui sono stati forniti dizionari tradotti in 4 lingue oltre a quello in italiano, mentre per le altre due variabili, Comune e Titolo di studio, è necessario gestire il quesito precodificato che precede la risposta la quesito al testo libero (la Provincia, nel caso del Comune e le 17 modalità per identificare il livello per il Titolo di studio). In entrambi i casi, devono essere gestite le casistiche relative alla presenza/assenza della risposta al quesito precodificato e alla coerenza/incoerenza dello stesso rispetto alla risposta testuale.

A tal fine, sono state fornite delle specifiche procedurali da seguire in fase di assegnazione del codice in modo da garantire il risultato rispetto ai criteri classificatori definiti in Istituto.

Al contrario, non sono stati forniti vincoli alla società in merito all'algoritmo di *matching* da utilizzare per individuare il codice, né sulle modalità in base alle quali ottenere il risultato, ossia in merito a quanto automatizzare e quanto lasciare ad un intervento manuale. Si è quindi preferito lasciare libera l'iniziativa della società, in quanto era possibile che già disponesse di pacchetti per il trattamento dei testi utilizzati per altre commesse; ci si è garantiti, però, pretendendo livelli minimi di assegnazione del codice e di accuratezza.

A tal fine sono stati impostati valori soglia per il tasso di codifica e quello di accuratezza analoghi a quelli del precedente censimento (Tavola 16).

Tavola 16 - Parametri per l'attività di codifica in *outsourcing*

VARIABILE	Livello minimo di assegnazione del codice	Livello minimo di accuratezza
Comune	95%	99%
Stato estero	90%	98%
Titolo di studio	85%	96%

Sono riportati nei paragrafi successivi i documenti forniti alla società incaricata di effettuare la codifica; tali documenti dovevano costituire una guida per il trattamento di ciascuna delle tre variabili; poiché nella fase di attribuzione dei codici è contemplata anche l'attività di un operatore manuale, essi forniscono non soltanto le specifiche procedurali, ma sono anche descrittivi dei contenuti delle basi informative.

6.1 Variabile Comune

6.1.1 Aspetti particolari da seguire per ottimizzare la codifica

Un primo fattore da tenere presente riguarda la variazione nel tempo di alcuni aspetti legati alle Province che possono avere un effetto sull'individuazione dei codici dei Comuni.

La **sigla della Provincia**, infatti, è di fondamentale importanza, rappresentando il filtro per la codifica dei testi. Per questo motivo è importante sapere che:

1. sono state introdotte due nuove sigle di provincia, ossia:
   - PU (Pesaro-Urbino) che ha sostituito la precedente PS (Pesaro);
   - FC (Forlì-Cesena) che ha sostituito la precedente FO (Forlì).

2. dal 1995 sono state istituite le seguenti 6 nuove province:
   - LC (Lecco) precedentemente appartenente a Como (CO) e Bergamo(BG);
   - LO (Lodi) precedentemente appartenente a Milano (MI);
   - VB (Verbano-Cusio-Ossola) precedentemente appartenente a Novara (NO);
   - BI (Biella) precedentemente appartenente a Vercelli (VC);
   - RN (Rimini) precedentemente appartenente alla provincia di Forlì-Cesena (FC);
   - PO (Prato) precedentemente appartenente a Firenze (FI);
   - KR (Crotone) precedentemente appartenente a Catanzaro (CZ);
   - VV (Vibo Valentia) precedentemente appartenente a Catanzaro (CZ).
3. dal 2006 sono state istituite le seguenti 4 nuove province:
   - OT (Olbia-Tempio) precedentemente appartenente a Sassari (SS) e Nuoro (NU);
   - OG (Ogliastra) precedentemente appartenente a Nuoro (NU);
   - VS (Medio Campidano) precedentemente appartenente a Cagliari (CA);
   - CI (Carbonia-Iglesias) precedentemente appartenente a Cagliari (CA).
4. dal 2009 sono state istituite le seguenti 3 nuove province:
   - MB (Monza e della Brianza) precedentemente appartenente a Milano (MI);
   - FM (Fermo) precedentemente appartenente ad Ascoli Piceno (AP);
   - BT (Barletta-Andria-Trani) precedentemente appartenente a Bari (BA) e Foggia (FG).

Per tutti i Comuni appartenenti a queste Province i rispondenti potrebbero indicare sia la vecchia che, alternativamente, la nuova sigla della Provincia. Per questo motivo tali Comuni sono stati inseriti, con la sigla attuale nel DIZIONARIO 1 e con la sigla della vecchia Provincia nel DIZIONARIO 2. Ovviamente il codice a 6 *digi*t (Comune + provincia) è lo stesso in entrambi i dizionari.
Inoltre, per alcuni dei Comuni ceduti all'estero è stata inserita una Provincia fittizia (es. ZA per Zara). Questo vale anche per i Rioni relativi a tali Comuni.

Un secondo fattore per ottimizzare la codifica riguarda il possibile trattamento dei testi nell'ottica di **standardizzazione** degli stessi.

È da tenere presente, innanzi tutto, che la base informativa utilizza un alfabeto che comprende anche i caratteri della lingua tedesca e i segni diacritici.

Inoltre, i testi sono scritti utilizzando gli Alti/Bassi, ma questo elemento, tuttavia, è ininfluente per l'attribuzione del codice; è opportuno quindi uniformare sia i testi da codificare che quelli della base informativa, riportando tutto al Maiuscolo o al Minuscolo.

Un terzo aspetto riguarda la gestione di **errori, abbreviazioni e modi di dire ricorrenti**.

Per ottimizzare l'efficienza del sistema di codifica automatica occorre considerare i vari modi di scrivere i Comuni, i modi più usati di abbreviare un nome e gli errori più frequenti. In particolare si è osservato quanto riportato nella tavola 17.

Tavola 17 - Errori, abbreviazioni e modi di dire ricorrenti

<table>
<tr>
<td>a) le abbreviazioni più frequenti sono:
• j. → jonico
• j.co → jonico
• j.ca → jonica
• s. → San,Santo, Santa
• v. → Val, Valle</td>
<td rowspan="3">d) I modi di scrivere più frequenti riguardano:
• La suddivisione in due parole di comuni costituiti da una parola composta. Sono tutti quei comuni contenenti le parole: monte, castel/castello, ponte, fiume, rio. Ad esempio:
FIUMEROSSO potrebbe essere scritto staccato
FIUME ROSSO
• La sostituzione della lettera J in I o della Y in I. Ad esempio:
BAGNAJA potrebbe essere scritta BAGNAIA
GRESSONEY potrebbe essere scritto GRESSONEI</td>
</tr>
<tr>
<td>b) gli errori più frequenti sono:
• proto → porto
• aqua → acqua</td>
</tr>
<tr>
<td>c) i sinonimi più frequenti sono:
• Lucania → Basilicata
• Puglie → Puglia
• Appula → Puglia</td>
</tr>
</table>

Un ultimo aspetto riguarda infine **le parole o costruzioni di testo ininfluenti,** infatti al fine di ottenere un maggior numero di codici univoci è consigliato di non utilizzare ai fini del match tutte quelle parole che sono inutili per individuare un Comune. Queste sono rappresentate da preposizioni (di, a da, con, dal , ecc.) e articoli (il, la, le, gli, ecc.).

Non considerare queste stringhe nel *matching* testuale fa sì che, ad esempio, il Comune "*MONTALBANO DI ELICONA*" venga individuato anche se scritto come "*MONTALBANO ELICONA*".

Un discorso analogo vale per **l'ordinamento delle parole**: un Comune il cui testo è composto da più di una parola è identico al Comune che contiene le stesse parole, ma con ordine diverso. Ciò è importante soprattutto se si considera che per alcuni Comuni il cambio di denominazione è a volte consistito solo nel modificare l'ordine delle parole; di conseguenza questi Comuni saranno riportati nel dizionario una sola volta e non due in quanto "nuova" e "vecchia" denominazione sono considerate uguali. La tavola seguente riporta alcuni esempi.

Tavola 18 - Esempi di denominazioni con ordine delle parole variato

ATTUALE DENOMINAZIONE	VECCHIA DENOMINAZIONE
Calvi San Nazzaro	San Nazzaro Calvi
Castrocaro e Terra del Sole	Terra del Sole e Castrocaro
Paderno' Robbiate	Robbiate Paderno'

Il processo di codifica, descritto nel sotto-paragrafo 6.1.2, termina con due possibili esiti:

- l'attribuzione di un codice univoco;
- la mancata individuazione di un codice con la conseguente non codifica del testo.

Qualora si presentasse questa ultima situazione, l'operatore manuale può valutare la possibilità di effettuare la ricerca del codice da assegnare sul dizionario degli **"Stati Esteri":** può avvenire, infatti, che il rispondente sbagli e specifichi lo Stato in cui è nato nel campo del "Comune". In caso di successo nell'individuazione del codice di uno Stato estero, questo dovrà comunque essere apposto nel campo dello Stato estero.

6.1.2 Descrizione del Processo di codifica

Il processo di codifica, schematizzato nel *flow-chart* della Figura 14, prevede due strade alternative a seconda del contenuto del **campo sigla provincia "PROV"** sul questionario che viene usato come **filtro:**

- se PROV contiene l'informazione, allora questa dovrà essere abbinata alla sigla riportata nel campo PROV dei dizionari e viene effettuata la **RICERCA CON IL FILTRO PROVINCIA, ossia soltanto tra i *record* della base informativa che contengono quella sigla nel campo PROV**;
- se PROV è vuoto, ossia il rispondente non ha indicato la sigla della provincia, allora viene effettuata la **RICERCA SENZA IL FILTRO PROVINCIA**.

RICERCA CON IL FILTRO PROVINCIA

Il campo della provincia contiene l'informazione richiesta: il testo da codificare, quindi, può essere ricercato non in tutto il dizionario, ma solo nella parte che contiene i Comuni la cui sigla della Provincia (colonna A) è uguale alla sigla riportata sul questionario. La procedura procede seguendo i seguenti passi:

1) la ricerca inizia nella sotto-sezione del **DIZIONARIO 1** e può:
   a) individuare un **codice univoco**, in tal caso lo attribuisce al testo e termina;
   b) non individuare alcun codice oppure trovare più di un codice: in tal caso prosegue ad analizzare il DIZIONARIO 2.

2) La **ricerca** all'interno del **DIZIONARIO 2** procede secondo i passi a) e b) sopra riportati. Se fallisce passa al DIZIONARIO 3.
3) Analoga procedura per il **DIZIONARIO 3**:
   a) se trova **un codice univoco**, lo attribuisce al testo e termina;
   b) se non trova il codice, passa alla RICERCA SENZA FILTRO.

RICERCA SENZA IL FILTRO PROVINCIA

In questo caso il testo da codificare viene ricercato all'interno di **tutta** la base informativa; la ricerca viene effettuata in due casi:

- quando il campo della sigla della Provincia in *input* è vuoto;
- oppure quando è fallita la ricerca con filtro. In questo caso, se la procedura termina con l'attribuzione di un codice univoco, nasce una discrepanza tra la sigla della Provincia PROV e il codice provincia assegnato. Questa discrepanza verrà risolta successivamente in Istat nella fase di controllo e correzione dei dati.

In entrambi i casi la procedura segue i seguenti passi (analoghi a quelli già descritti),ossia:

1) la **ricerca** inizia all'interno di tutto il **DIZIONARIO 1** e può:
   a) individuare un **codice univoco**, in tal caso lo attribuisce al testo e termina;
   b) non individuare alcun codice oppure trovare più di un codice: in tal caso prosegue ad analizzare il **DIZIONARIO 2**.
2) La **ricerca** continua all'interno del **DIZIONARIO 2** secondo i passi a) e b). Se fallisce passa al DIZIONARIO 3.
3) Analoga procedura per il **DIZIONARIO 3**:
   a) se trova un **codice univoco**, lo attribuisce al testo e termina;
   b) se non trova il codice, allora richiede l'intervento dell'operatore manuale.

INTERVENTO DELL'OPERATORE MANUALE

L'intervento dell'operatore manuale si verifica sia nel caso di assenza della sigla della Provincia nel questionario che di presenza della stessa; nel primo caso anche l'operatore manuale effettuerà la RICERCA SENZA FILTRO, nel secondo, analogamente a quanto previsto per il *batch*, effettuerà dapprima la RICERCA CON FILTRO e, se questa non da adito ad un codice univoco, la RICERCA SENZA FILTRO.

Nel caso della RICERCA CON FILTRO, la procedura dovrà proporre dapprima all'operatore solo i comuni appartenenti alla provincia di input fra i quali ricercare un eventuale codice univoco da assegnare al testo e soltanto successivamente dargli la possibilità di allargare la ricerca in tutto il dizionario.

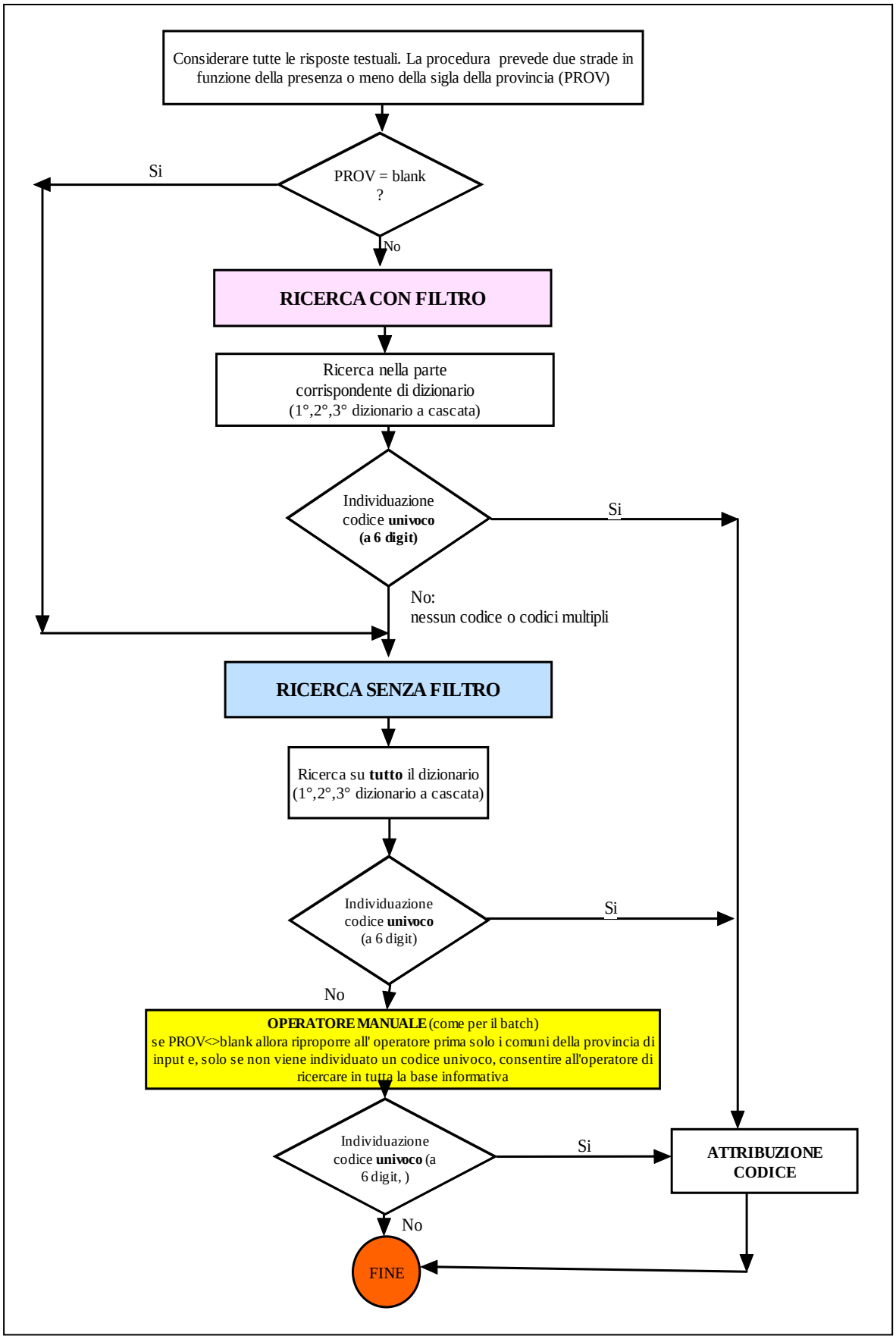
L'intervento di un operatore manuale è necessario ai fini della codifica in quanto questi potrebbe capire, dalla formulazione della risposta di input, quale sia il Comune di appartenenza; il testo potrebbe riportare, infatti, una dicitura non prevista nel dizionario, e quindi non codificabile automaticamente, ma tale da dare un'indicazione chiara a chi la legge. Ad es.: se il testo di input è "*San Georgeo (NA)*" la procedura automatica non troverà il Comune, essendo errato il nome e non esistendo Comuni con la sigla della Provincia inclusa nella denominazione, mentre l'operatore manuale potrà capire che si tratta di un Comune in provincia di Napoli la cui denominazione è "San Giorgio".

L'operatore potrebbe quindi:

a) trovare un **codice univoco**, attribuirlo al testo e terminare la procedura;
b) non trovare il codice: la procedura fallisce e il testo non potrà essere codificato.

È necessario porre attenzione al fatto che nel caso l'operatore effettui la ricerca senza filtro, ma la descrizione del Comune da codificare sia ambigua, in quanto l'operatore trova nella base informativa più Comuni con il nome simile appartenenti a Province diverse, deve assolutamente astenersi dall'attribuzione del codice.

Figura 14 - Procedura per la codifica del Comune



6.2 Variabile Titolo di studio

Anche per la classificazione del Titolo di studio, è stato necessario fornire specifiche dettagliate per la società incaricata di effettuare la codifica in modo tale da delineare alcuni aspetti procedurali e, nel contempo, circoscrivere la discrezionalità nell'interpretazione dei testi.

A tal fine sono stati forniti alcuni elementi funzionali a garantire l'ottimizzazione del processo di attribuzione del codice, come riportati nel sotto-paragrafo 6.2.1.

Si riporta invece nel sotto-paragrafo 6.2.2. la descrizione puntuale del processo di codifica da eseguire.

6.2.1 Aspetti particolari da seguire per ottimizzare la codifica

Ai fini dell'abbinamento delle stringhe alfabetiche digitate con le descrizioni presenti nel *file*/dizionario fornito, si esplicitano le stringhe che **non risultano significative**, ai fini del *match*:

- Gli articoli come per es.: *LA, LE, LO, GLI, IL, …..*;
- Le preposizioni come per es.: *DI, A, DA, IN, CON …..*;
- Le congiunzioni come per es.: *E, ED, …*;
- I suffissi finali. L'eliminazione di questi ultimi, infatti, permette di non differenziare tra maschile/femminile e singolare/plurale. Quelli ininfluenti sono: *A, E, I, HI, HE, IA, IE, IO, O, RICE, ICE, ORE, ORI, ERE, ERA*. Pertanto, chiarendo con un esempio, il Titolo di studio di *Ragioniere* è esattamente lo stesso di quello di *Ragioniera,* a meno del suffisso finale, e nel dizionario è presente solo al maschile singolare.

Nel dizionario non sono state usate le abbreviazioni che, invece, possono talvolta essere usate dai rispondenti. Di seguito (Tavola 19) si elencano quelle più utilizzate, sulla base delle esperienze finora effettuate.

Tavola 19 - Abbreviazioni frequentemente utilizzate

ABBREVIAZIONI	TRASFORMAZIONE
ABIL	ABILITAZIONE
ABILITAZ	ABILITAZIONE
ABILIT	ABILITAZIONE
ADD	ADDETTO
AMM	AMMINISTRATIVO
ASS	ASSISTENTE
ASSIST	ASSISTENTE
ATTEST	ATTESTATO
AZ - AZ.LE -	AZIENDALE
COMM – COMM.LE	COMMERCIALE
COM	COMMERCIALE
CORR	CORRISPONDENTE
DIPL	DIPLOMA
DIP	DIPLOMA
LIC.SC.FICO	LICEO SCIENTIFICO
MATUR	DIPLOMA
ELEM	ELEMENTARE
ELET	ELETTRICHE
ESP	ESPERTO
IND	INDUSTRIALE
INF	INFORMATICA
IST	ISTITUTO
I	PRIMO
II	SECONDO
MASTER	LAUREA
MECC	MECCANICO
OP	OPERATORE

Tavola 19 segue **- Abbreviazioni frequentemente utilizzate**

ABBREVIAZIONI	TRASFORMAZIONE
OPER	OPERATORE
PER	PERITO
PROF - PROF./LE	PROFESSIONALE
PROFESS PROF.LE	PROFESSIONALE
PROGR	PROGRAMMATORE
QUALIF	QUALIFICA
RAG	RAGIONIERE
SEG	SEGRETERIA
SEGR	SEGRETERIA
SPEC	SPECIALIZZATO
STAT	STATALE
SUP	SUPERIORE
TEC	TECNICO
TECN	TECNICO
TECNOLOL	TECNOLOGIE
TELEC	TELECOMUNICAZIONE

Nella fase di *matching* è bene, qualora le abbreviazioni usate siano di facile interpretazione e riconducibili univocamente ad una sola parola, esplicitarle per consentire una più agevole ricerca all'interno del dizionario.

6.2.2 Descrizione del Processo di codifica

È necessario premettere che nella fase di ricerca per la codifica del Titolo di studio si devono considerare tutte le risposte testuali, fornite nel Quesito a Testo Libero, indipendentemente se dovute o meno sulla base del Quesito Precodificato (QP).

Sulla base del *flow chart* di seguito riportato (Figura 15), la procedura di codifica può seguire i seguenti due iter.

A. Nel questionario è **presente la risposta al QP**;

B. Nel questionario **non è presente la risposta al QP,** oppure **tale risposta non ha consentito l'individuazione di un codice univoco**.

Procediamo ad illustrare ciascuno dei due citati iter.

A. NEL QUESTIONARIO È PRESENTE LA RISPOSTA AL QP.

La ricerca deve essere effettuata soltanto nella sezione di dizionario in cui il valore del campo filtro corrisponde alla risposta fornita al QP (**ricerca con filtro**).

Se presenti più biffature, ai fini della ricerca con filtro, deve essere utilizzata quella con valore più elevato, qualora non si individui il Titolo di studio nell'ambito del filtro con valore più elevato, il titolo deve essere ricercato nel sotto-insieme di dizionario corrispondente al filtro con valore inferiore, prima di passare alla ricerca senza filtro (cfr. **Punto B**).

Si possono verificare tre casi.

A.1) L'esito della ricerca sulla base del testo fornito dal rispondente **consente l'individuazione di un codice univoco.** Gli esempi di seguito riportati individuano i due casi nei quali la selezione implica l'estrazione di un singolo *record* dal dizionario oppure di più *record* cui corrisponde sempre lo stesso codice.

Per esempio: risposta testuale = ***'avifauna'*** e biffatura al QP della modalità 06 *(Diploma di istituto professionale)* con la specifica della durata del corso pari a 2-3 anni.

Tavola 20 - Esempio di individuazione di codice univoco

FILTRO	Codice	F	Titolo di studio
061	30101		Avifauna

Come si evince dalla tavola 20, il Titolo di studio è univocamente codificabile, essendo compatibile con quanto specificato nel QP.

Ulteriore esempio: risposta testuale = ***'diploma orafo'*** e biffatura al QP della modalità 06 *(Diploma di istituto professionale)* con la specifica della durata del corso pari a 2-3 anni (Tavola 21)

Tavola 21 - Ulteriore esempio di individuazione di codice univoco

FILTRO	Codice	F	Titolo di studio
061	30102		Operatore meccanico orafo
061	30102		Operatore orafo
061	30102		Orafo

Avremmo analogamente l'assegnazione di un codice univoco sebbene dal *match* testuale siano stati individuati nel dizionario più descrizioni, tutte corrispondenti allo stesso codice.

A.2) l'esito della ricerca sulla base del testo fornito dal rispondente **non consente l'individuazione di un codice univoco,** in quanto vengono estratti dal dizionario più *record* cui sono associati codici diversi.

È necessario pertanto procedere alla ricerca così come descritto al **punto B)**.

Per esempio: risposta testuale = ***'diploma di perito'*** e biffatura al QP della modalità 09 *(Diploma di istituto tecnico).*

Tavola 22 - Risultato della query 'diploma di perito'

FILTRO	Codice	F	Titolo di studio
090	40201		DIPLOMA PERITO AGRARIO
090	40202		DIPLOMA PERITO ELETTROTECNICO SPECIALISTA ELETTRONICA INDUSTRIALE
090	40202		DIPLOMA PERITO ELETTROTECNICO
090	40202		DIPLOMA PERITO TERMOTECNICA
090	40202		DIPLOMA PERITO INDUSTRALE
090	40204		DIPLOMA PERITO AEREONAUTICO

In questo caso, come si evince dalla tavola 22, il codice non è univoco (ultimo digit è diverso); è necessario pertanto procedere alla ricerca così come descritto al punto B).

A.3) L'esito della ricerca sulla base del testo fornito dal rispondente **non consente l'individuazione di alcun codice.** Fallisce cioè l'abbinamento tra la risposta testuale e i testi presenti nel dizionario.

In tal caso si procede alla ricerca descritta come al **punto B)**.

È necessario porre attenzione al fatto che è possibile la presenza nel questionario di una risposta parziale al QP (non esplicitazione della durata per le modalità di risposta 06, 07 e 08)

È possibile infatti che il rispondente abbia risposto ad una delle 17 modalità sul Titolo di studio, ma non al quesito 5.4 sulla durata (prevista per le modalità di risposta 06, 07 e 08).

In tal caso, la ricerca deve essere effettuata nel sotto-insieme di dizionario i cui primi due *digit* del filtro corrispondano alla modalità di risposta selezionata, per verificare se sia individuabile un titolo di studio univocamente attribuibile.

Soltanto nel caso in cui non si individui un codice univoco, si dovrà passare alla ricerca senza filtro (cfr. **Punto B**).

Per esempio: è' stato biffato *06 – Istituto professionale*, senza specificare la durata del corso di studio ed è stata fornita una risposta testuale: il *matching* sarà effettuato nel sotto-insieme di dizionario contenente i diplomi di istituto professionale (primi due digiti del filtro = 06). Il codice sarà attribuito soltanto nel caso in cui si individui un unico Titolo di studio corrispondente al testo digitato; qualora invece non se ne individui nessuno o se ne individuino due (o più), si dovrà passare alla ricerca senza filtro (cfr. **punto B**).

Qualora invece sia stato risposto al quesito 5.4 quando non dovuto (non in corrispondenza delle modalità di risposta 06, 07 e 08), la biffatura fornita al quesito 5.4 non dovrà essere considerata ai fini dell'attribuzione del codice.

B. NEL QUESTIONARIO **NON È PRESENTE LA RISPOSTA AL QP**, OPPURE **TALE RISPOSTA NON HA CONSENTITO L'INDIVIDUAZIONE DI UN CODICE UNIVOCO** (vedi punti A.2 e A.3).

La ricerca ora deve essere effettuata su tutto il dizionario (ricerca senza filtro).

Si possono verificare tre casi.

B.1) L'esito della ricerca sulla base del testo fornito dal rispondente **consente l'individuazione di un codice univoco.**

Per esempio: risposta testuale = '***agrotecnico'*** (Tavola 23)

Tavola 23 - Esempio di individuazione di codice univoco

FILTRO	Codice	F	Titolo di studio
062	40101		Agrotecnico

In questo caso è possibile l'attribuzione **univoca del codice** poiché il Titolo di studio dichiarato proviene esclusivamente da un diploma di istituto professionale della durata di 4-5 anni, quindi è identificabile anche in caso di biffatura errata o mancante.

B.2) L'esito della ricerca sulla base del testo fornito dal rispondente **non consente l'individuazione di un codice univoco, ma di un set di possibili codici.**

In tal caso si verifica se, tra il set di possibili codici, è soddisfatta la seguente condizione:

Uno ed uno solo inizia per 9 (= codice generico) e non ha filtro (blank nella colonna A)

Se la condizione è verificata si assegna il codice che inizia per 9, infatti, la risposta testuale non ha un contenuto informativo tale da identificare univocamente un Titolo di studio, ma corrisponde ad un insieme di titoli che sono però associati univocamente ad un codice generico.
Per esempio: risposta testuale = '***arti grafiche'*** (Tavola 24)**.**

Tavola 24 Risultato della *query* 'arti grafiche'

Filtro	Codice	F	Titolo di studio
081	30601		Arti grafiche
090	40202		Arti grafiche
090	40202		Arti grafiche editoriali e pubblicitarie
090	40202		PROGETTO "TEMPT" - ARTI GRAFICHE
082	40601		Arti grafiche
	93040		Arti grafiche

Se invece questa condizione non si verifica dovrà subentrare **l'operatore manuale**, che verificherà:

- se è possibile attribuire un codice;
- se non è possibile attribuire alcun codice.

B.3) L'esito della ricerca sulla base del testo fornito dal rispondente **non consente l'individuazione di alcun codice.** Fallisce cioè l'abbinamento tra la risposta testuale e i testi presenti nel dizionario.

Anche in questo caso dovrà subentrare **l'operatore manuale** che verificherà:

- se è possibile attribuire un codice;
- se non è possibile attribuire alcun codice.

Si specifica che, qualora si giunga a questo *step* (all'operatore manuale) provenendo da **A.2)** (era stato individuato un set di codici possibili nella sottosezione di dizionario in cui il valore del campo filtro corrisponde alla risposta fornita al QP), saranno riproposti all'operatore manuale soltanto i testi del dizionario corrispondenti alla biffatura al QP. L'operatore in questo caso potrà:

- attribuire un codice univoco generico (ultimi due *digit* del codice uguali a 00), vedi il caso dell'esempio risposta testuale = ***'Diploma istituto tecnico'.*** In questo caso, infatti, non essendo state fornite le informazioni di dettaglio non è possibile individuare il tipo di Istituto tecnico frequentato;
- attribuire un codice univoco dettagliato. Se fosse stata rilevata una risposta testuale con un errore ortografico come per esempio ***'perito agario',*** il sistema automatico potrebbe aver fallito il *matching*, mentre l'operatore potrebbe facilmente attribuire il codice dettagliato del ***'perito agrario***';
- non attribuire alcun codice perché la risposta non è riconducibile ad alcun Titolo di studio.

Si riporta un esempio di risposta **non** codificabile in caso di **assenza della biffatura** al QP, quale risposta testuale = ***'magistrale'***.

La ricerca nel dizionario evidenzia quanto riportato nella tavola sottostante.

Tavola 25 - Risultato della *query* "magistrale"

FILTRO	Codice	F	Titolo di studio
071	30301		SCUOLA MAGISTRALE (3 anni)
071	30301		MAGISTRALE DI GRADO PREPARATORIO
072	40301		Diploma di istruzione secondaria superiore di Scuola magistrale
072	40301		Scuola magistrale
072	40301		SCUOLA MAGISTRALE (5 anni)
100	40302		Abilitazione magistrale
100	40302		Diploma di istruzione secondaria superiore di Istituto magistrale
100	40302		DIPLOMA DI MATURITA' (DI STATO) MAGISTRALE
100	40302		Istituto magistrale quinquennale (incluso anno integrativo)
100	40302		Magistrale
100	40302		ANNO INTEGRAZIONE MAGISTRALE
100	40302		DIPLOMA MAGISTRALE SPERIMENTALE LINGUISTICO
100	40302		MATURITA' MAGISTRALE SPERIMENTALE LINGUISTICO
170	73000000		Laurea magistrale a ciclo unico

In questo caso l'operatore, mancando in origine la biffatura al QP, non è in grado di attribuire nessun codice.

INTERVENTO DELL'OPERATORE MANUALE

Come già detto, il codificatore manuale, che subentra soltanto nei casi in cui la procedura automatica non è riuscita ad individuare un codice univoco, può trovarsi in due situazioni:

1. può effettuare la ricerca in tutto il dizionario;
2. è limitato nella ricerca in una sottosezione di dizionario se nel questionario era presente la risposta al quesito QP e si proveniva dal passo **A.2)** (era stato individuato un set di codici possibili nella sottosezione di dizionario in cui il valore del campo filtro corrispondeva alla risposta fornita al QP).

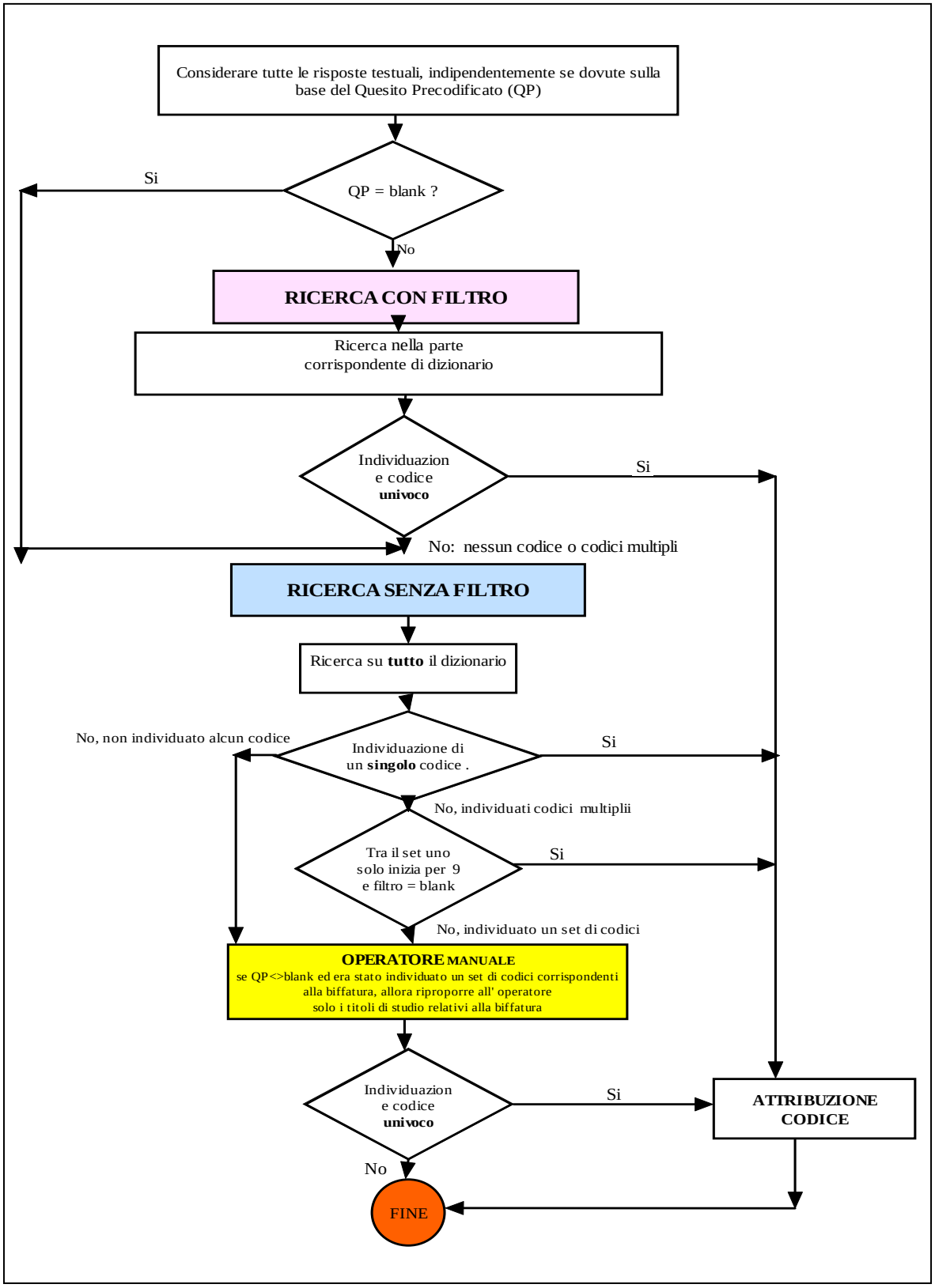
Si vuole qui fornire alcune indicazioni che possono aiutare il codificatore qualora si trovi nella prima delle due situazioni descritte, ossia debba ricercare il codice in tutto in dizionario in quanto nel questionario non era presente la risposta al quesito QP oppure, anche se presente, non era stato individuato un set di codici possibili nella sottosezione di dizionario coerente con la biffatura.

Si segnala che in questi casi alcune parole possono guidare il codificatore nella sottosezione di dizionario in cui ricercare il codice.
Si riportano nella seguente tavola le parole che possono avere questa funzione di 'guida'.

Tavola 26 - Parole con funzioni di 'guida' nell'individuazione del codice

PAROLE/STRINGHE	Sezioni di dizionario corrispondenti ai relativi filtri
Diploma di laurea, laurea …	140, 150, 160, 170
Diploma universitario	120, 130
Laurea breve, minilaurea	120, 130, 140,150
Diploma accademico di I livello	140, 150
Diploma accademico di II livello	160, 170
Laurea I livello, laurea triennale, laurea triennale nuovo ordinamento	150
Laurea II livello, laurea specialistica, laurea magistrale, laurea a ciclo unico, laurea specialistica biennale, laurea vecchio ordinamento, laurea secondo livello	170, 160
Qualifica, diploma di qualifica	050, 061, 071, 081
Diploma, diploma di istruzione secondaria superiore, maturità, diploma di stato, diploma di scuola secondaria superiore di secondo grado	062, 072, 082, 090, 100, 110

Figura 15 - Procedura per la codifica del Titolo di Studio



6.3 Variabile Stato estero

6.3.1 Aspetti particolari da seguire per ottimizzare la codifica

Per ottimizzare l'efficienza del sistema di codifica automatica, occorre considerare:

abbreviazioni frequentemente utilizzate dai rispondenti, quali:

- rep. → repubblica;
- dem → democratica;
- pop. → popolare;
- fed. → federale, federativo, federazione.

modi di scrivere nomi composti, quali:

- Capo Verde = Capoverde;
- Costa Rica = Costarica.

6.3.2 Descrizione del Processo di codifica

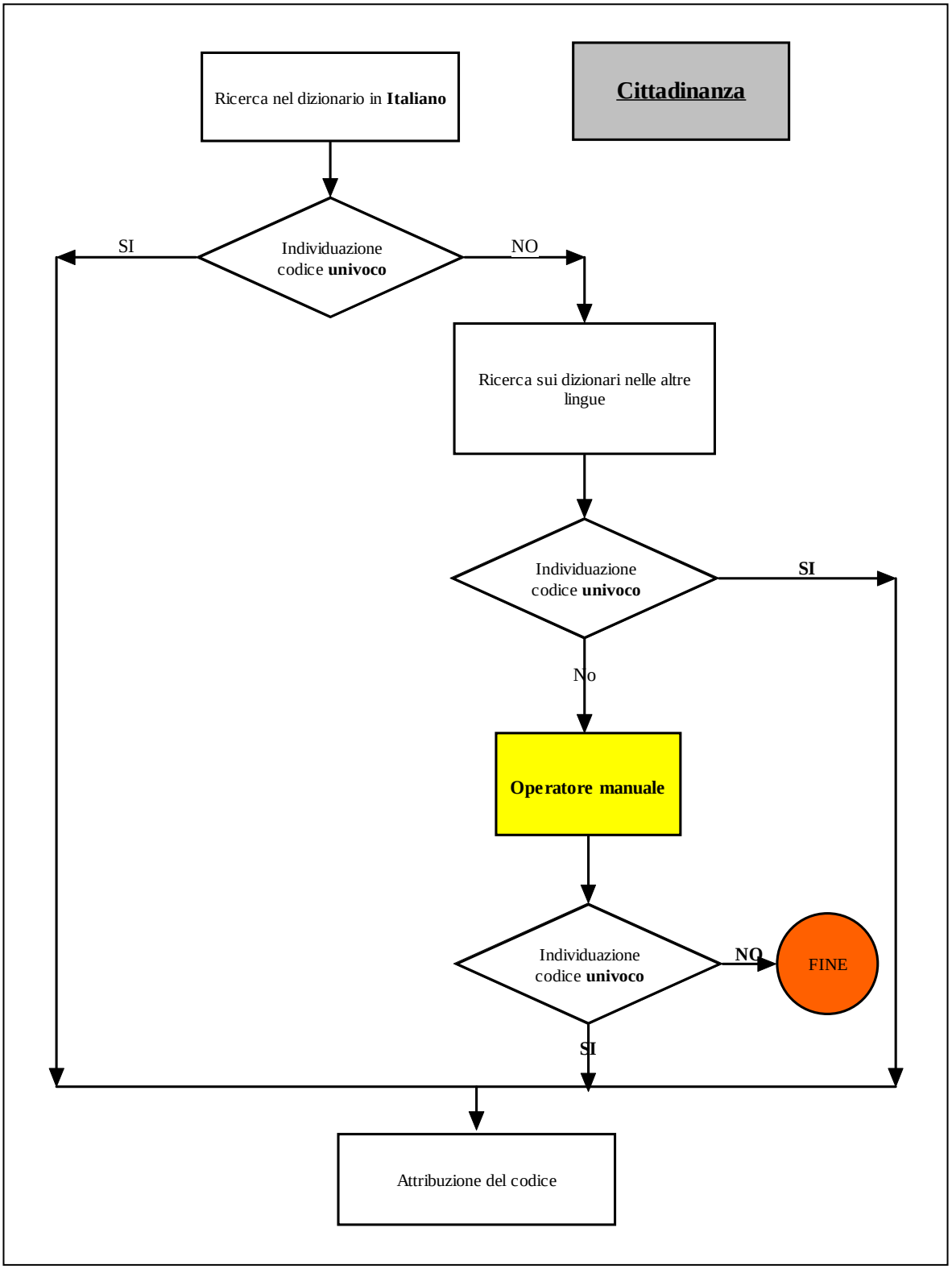
La procedura ottimale per l'attribuzione del codice alla stringa da codificare è quella di effettuare la ricerca nel dizionario in Italiano e, soltanto qualora non si riesca ad individuare il codice, di ricercare nei dizionari in lingua straniera (Figura 16).

Le esperienze finora effettuate hanno infatti dimostrato che gli intervistati tendono a rispondere in Italiano e, soltanto raramente, si esprimono in un'altra lingua.

Qualora anche tale ricerca dia esito negativo, il caso deve essere sottoposto al codificatore manuale, che determinerà se la risposta ha un contenuto informativo sufficiente per essere univocamente codificata, oppure no.

Qualora a seguito delle operazioni descritte non si riuscisse ad attribuire il codice, l'operatore manuale dovrà effettuare la ricerca del codice da assegnare anche sul dizionario dei "**Comuni":** può avvenire, infatti, che il rispondente, per errore, specifichi il nome del Comune nel campo riservato allo "Stato estero". In caso di successo nell'individuazione del codice di un Comune, questo dovrà essere attribuito e apposto nel campo destinato al codice del Comune.

Figura 16 - Procedura per la codifica della variabile Stato estero/Cittadinanza



7. Il Navigatore delle Professioni

7.1 La base informativa del navigatore delle Professioni

Il valore informativo della variabile 'Professione' è stato riconosciuto fin dal primo Censimento generale del Regno d'Italia, quando fu reso obbligatorio per ogni cittadino dichiarare la propria condizione lavorativa. Erano note già da allora le difficoltà connesse alla rilevazione del lavoro svolto e il conseguente impegno richiesto per garantire standard elevati di qualità ai dati statistici sulle professioni.

Dal punto di vista metodologico, infatti, codificare l'attività lavorativa di un occupato all'interno della tassonomia ufficiale delle Professioni comporta uno sforzo concettuale non irrilevante. In primo luogo ciò è dovuto al rapporto tra la numerosità dei mestieri esistenti - se ne contano più di 20 mila – e la consistenza del livello di aggregazione di minore dettaglio entro cui è possibile ricondurli.[15] un rapporto che determina una elevata eterogeneità tra le Professioni ricomprese all'interno degli stessi raggruppamenti e riduce inevitabilmente la specificità delle definizioni indicate per orientare l'attività di codifica.

In secondo luogo, è da tenere presente che i confini tra i diversi raggruppamenti non sono nettamente definiti e si prestano a continue sovrapposizioni. Inoltre, nomi di Professioni simili possono sottendere attività lavorative diametralmente opposte, come pure nomi differenti richiamare insiemi di mansioni fondamentalmente simili.

Non per ultimo la molteplicità dei punti di vista attraverso cui è possibile rappresentare il lavoro ingenera delle indebite sovrapposizioni concettuali minando l'omogeneità dei dati raccolti. Nel classificare la propria Professione, ad esempio, è frequente confondere le diverse prospettive di lettura del fenomeno, orientando la scelta del codice in funzione di elementi come l'Attività economica, la posizione nella Professione o l'inquadramento contrattuale che non risultano informativi dell'attività svolta.

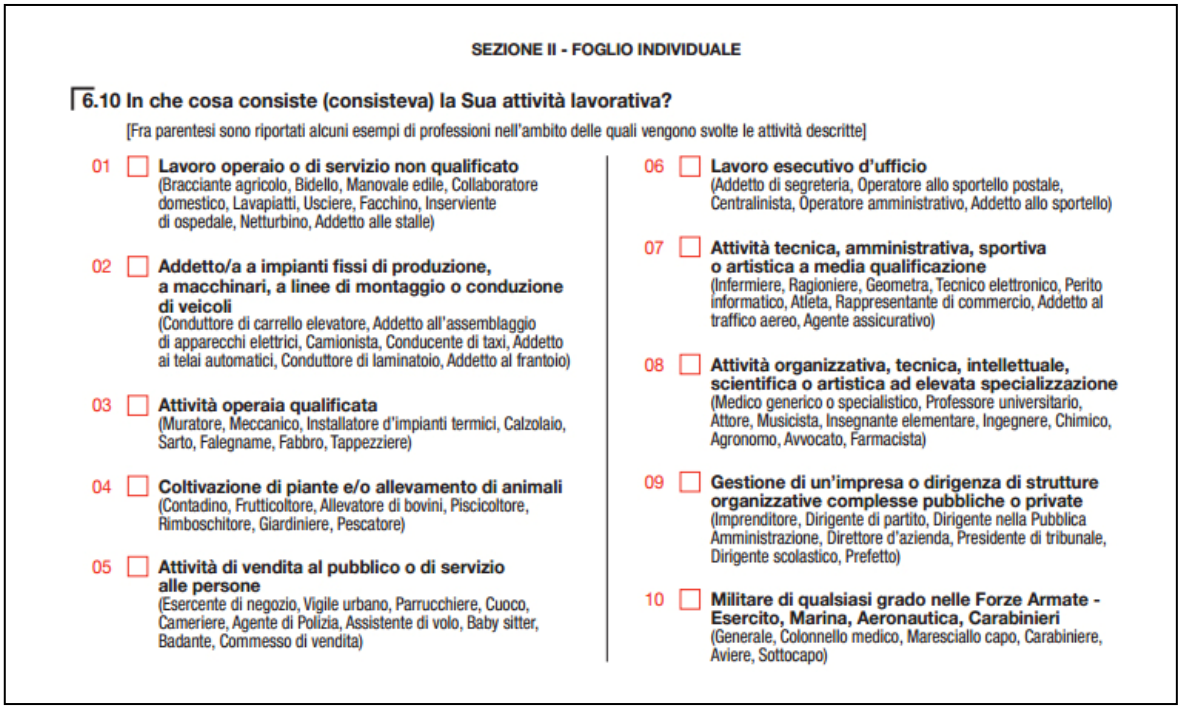
Per ridurre le difficoltà appena elencate e orientare in maniera omogenea le procedure di codifica, la classificazione delle Professioni è fondata su un criterio univoco e oggettivamente ravvisabile di rappresentazione del lavoro: la tipologia di mansioni svolte. La definizione sottesa da questa impostazione definisce la Professione "*come un complesso di attività lavorative concrete, unitarie rispetto all'individuo che le svolge, che richiama, a vari livelli, statuti, conoscenze, competenze, identità e sistemi di relazione propri*".[16]

Nel progettare la rilevazione della Professione nell'ultimo censimento della popolazione si è partiti da questa definizione, indirizzando l'attività di codifica sul riconoscimento dei compiti svolti dai cittadini nell'esercizio della loro occupazione. Nello specifico, il richiamo alla concretezza delle mansioni eseguite è stato declinato nella formulazione di un quesito (Figura 17) che focalizzasse l'attenzione dei rispondenti sulle finalità delle loro attività quotidiane, a prescindere dalla considerazione di altri elementi che connotano tradizionalmente il lavoro.

[15] La versione attuale della Isco08 (International Standard Classification of Occupation) prevede 10 grandi gruppi al primo livello della classificazione.

[16] Istat, Classificazione delle professioni, Metodi e norme, Nuova serie - N.12, p.16, Roma, 2001.

Figura 17 - Quesito per rilevare la Professione svolta nel modello censuario 2011



SEZIONE II - FOGLIO INDIVIDUALE

6.10 In che cosa consiste (consisteva) la Sua attività lavorativa?

[Fra parentesi sono riportati alcuni esempi di professioni nell'ambito delle quali vengono svolte le attività descritte]

01 ☐ **Lavoro operaio o di servizio non qualificato**
(Bracciante agricolo, Bidello, Manovale edile, Collaboratore domestico, Lavapiatti, Usciere, Facchino, Inserviente di ospedale, Netturbino, Addetto alle stalle)

02 ☐ **Addetto/a a impianti fissi di produzione, a macchinari, a linee di montaggio o conduzione di veicoli**
(Conduttore di carrello elevatore, Addetto all'assemblaggio di apparecchi elettrici, Camionista, Conducente di taxi, Addetto ai telai automatici, Conduttore di laminatoio, Addetto al frantoio)

03 ☐ **Attività operaia qualificata**
(Muratore, Meccanico, Installatore d'impianti termici, Calzolaio, Sarto, Falegname, Fabbro, Tappezziere)

04 ☐ **Coltivazione di piante e/o allevamento di animali**
(Contadino, Frutticoltore, Allevatore di bovini, Piscicoltore, Rimboschitore, Giardiniere, Pescatore)

05 ☐ **Attività di vendita al pubblico o di servizio alle persone**
(Esercente di negozio, Vigile urbano, Parrucchiere, Cuoco, Cameriere, Agente di Polizia, Assistente di volo, Baby sitter, Badante, Commesso di vendita)

06 ☐ **Lavoro esecutivo d'ufficio**
(Addetto di segreteria, Operatore allo sportello postale, Centralinista, Operatore amministrativo, Addetto allo sportello)

07 ☐ **Attività tecnica, amministrativa, sportiva o artistica a media qualificazione**
(Infermiere, Ragioniere, Geometra, Tecnico elettronico, Perito informatico, Atleta, Rappresentante di commercio, Addetto al traffico aereo, Agente assicurativo)

08 ☐ **Attività organizzativa, tecnica, intellettuale, scientifica o artistica ad elevata specializzazione**
(Medico generico o specialistico, Professore universitario, Attore, Musicista, Insegnante elementare, Ingegnere, Chimico, Agronomo, Avvocato, Farmacista)

09 ☐ **Gestione di un'impresa o dirigenza di strutture organizzative complesse pubbliche o private**
(Imprenditore, Dirigente di partito, Dirigente nella Pubblica Amministrazione, Direttore d'azienda, Presidente di tribunale, Dirigente scolastico, Prefetto)

10 ☐ **Militare di qualsiasi grado nelle Forze Armate - Esercito, Marina, Aeronautica, Carabinieri**
(Generale, Colonnello medico, Maresciallo capo, Carabiniere, Aviere, Sottocapo)

Come nel 2001 e analogamente a quanto stabilito per il quesito sull'Attività economica, per il Censimento del 2011 i rispondenti sono stati chiamati ad auto codificarsi ovvero a scegliere quale delle dieci modalità proposte rappresentasse al meglio la propria attività lavorativa. L'attenzione posta in sede di progettazione del questionario è stata quindi focalizzata sul *wording* del quesito 6.10 nel tentativo di sintetizzare con un linguaggio univoco e intellegibile da tutta la popolazione le tipologie di attività eseguite per ciascuno dei raggruppamenti previsti.

A fianco delle definizioni proposte si è ritenuto opportuno inserire delle esemplificazioni di nomi di professioni contenute nei diversi raggruppamenti, così da affiancare all'inevitabile genericità delle descrizioni dei riferimenti concreti a lavori ritenuti rappresentativi delle modalità di risposta, nonché sufficientemente noti alla popolazione italiana. L'insieme delle informazioni proposte ha cercato in questo modo di ridurre l'inevitabile smarrimento che si registra in presenza di un linguaggio classificatorio non sempre in linea con il linguaggio del senso comune.

Considerata l'importanza dell'operazione censuaria e tenuto conto delle difficoltà intrinseche nell'attività di codifica che sono state appena richiamate, si è inteso predisporre uno strumento aggiuntivo di ausilio alla compilazione del quesito, il Navigatore a supporto del quesito censuario sulle Professioni, che facilitasse i cittadini nell'individuare la risposta più adatta a rappresentare il proprio lavoro.

Tale navigatore, le cui funzioni e i criteri di ricerca saranno descritti nel paragrafo 7.3, riprende a sua volta, adattandolo alle esigenze censuarie, lo strumento di consultazione della Classificazione ufficiale delle Professioni CP2011, illustrato nel paragrafo 7.2.

7.2 Il navigatore della classificazione delle Professioni CP2011

Il 1° gennaio 2011 è entrata ufficialmente in vigore la nuova Classificazione italiana delle professioni (CP2011), frutto di un lavoro di aggiornamento della precedente versione (CP2001) e di adattamento alle novità introdotte dalla Classificazione internazionale Isco08.

La classificazione italiana aderisce pienamente all'impostazione prescelta a livello internazionale, condividendone i criteri generali ma preservando le specificità proprie del mercato del lavoro italiano.

Così come per la Classificazione internazionale Isco08, è il criterio della competenza, sopra richiamato, a delineare il sistema classificatorio della CP2011, articolato su 5 livelli di aggregazione gerarchici:

- il primo livello, di massima sintesi, composto da 9 grandi gruppi professionali;
- il secondo livello, comprensivo di 37 gruppi professionali;
- il terzo livello, con 129 classi professionali;
- il quarto livello, formato da 511 categorie;
- il quinto e ultimo livello della classificazione, con 800 unità professionali, dentro cui sono riconducibili le professioni esistenti nel mercato del lavoro.

La classificazione propone inoltre, per ciascuna unità professionale, un elenco di voci professionali, che viene riportato a titolo esemplificativo per orientare e facilitare il lettore nella consultazione e nella ricerca.

Sebbene la Classificazione italiana sia pienamente raccordabile alla Classificazione internazionale Isco08, non vi è una corrispondenza biunivoca fra i primi livelli gerarchici delle due classificazioni: il primo livello della Isco08, infatti, è composto da 10 raggruppamenti professionali mentre la CP2011 ne conta 9. Le tabelle di transcodifica ufficiali prevedono una ricostruzione del terzo livello della Isco08 attraverso il quinto livello della classificazione italiana CP2011.

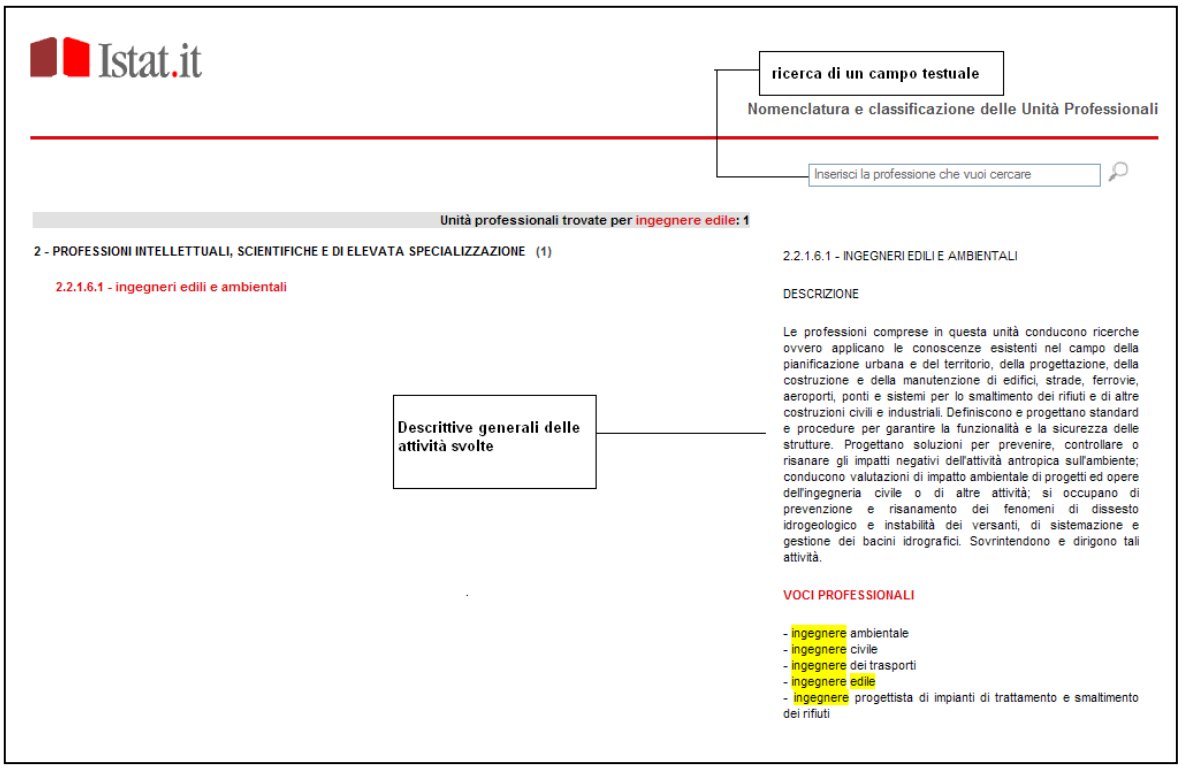
Per agevolare la consultazione della nuova Classificazione delle Professioni è stato predisposto il 'navigatore della Classificazione delle Professioni'.

Tale applicativo, disponibile nel sito dell'Istat all'indirizzo http://cp2011.istat.it/, permette di individuare la collocazione di una Professione all'interno della gerarchia prevista dalla Classificazione italiana CP2011 seguendo due modalità alternative, che prevedono:

- la ricerca di un campo testuale;
- la scelta del livello classificatorio via, via più dettagliato, guidata dal nome e dalle descrittive generali delle attività lavorative che contraddistinguono i diversi raggruppamenti professionali.

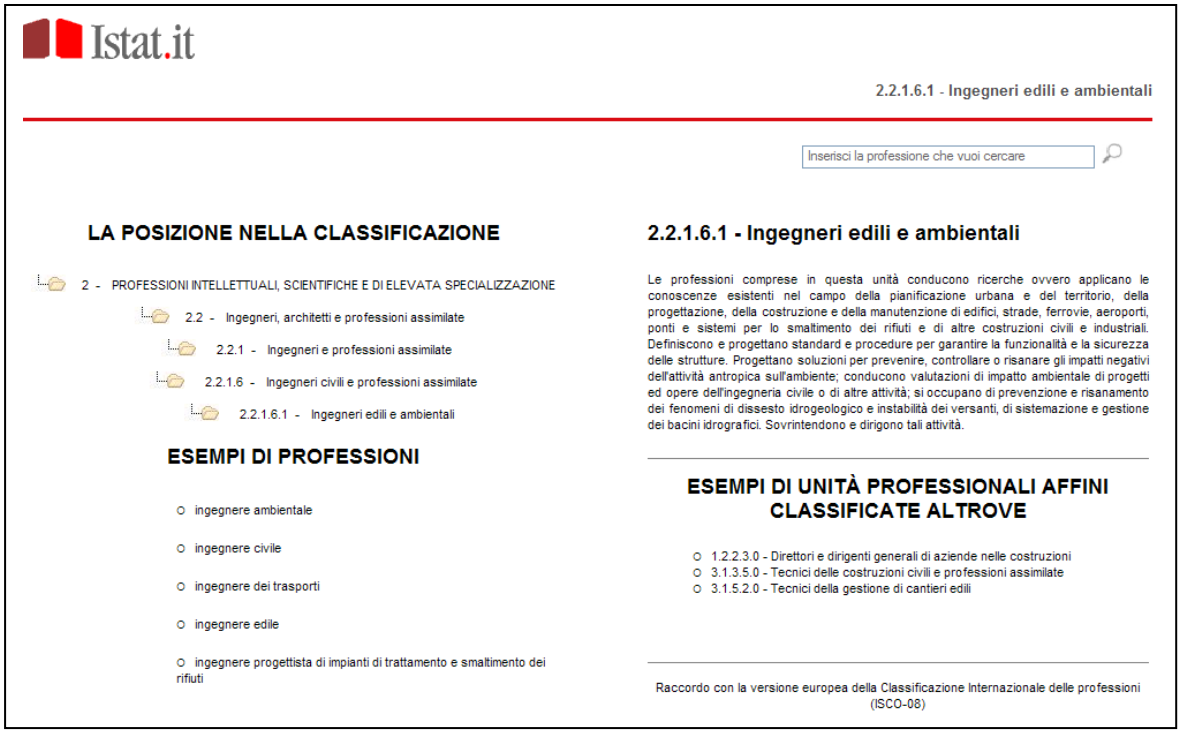
Così ad esempio, la ricerca della stringa testuale 'ingegnere edile' porterà alla risposta riportata in figura 18.

Figura 18 - Output della ricerca nel navigatore della CP2011 della Professione di "ingegnere edile"



Alternativamente, il ricorso alla navigazione gerarchica della classificazione impone una scelta in successione, coadiuvata dalle informazioni sulle attività lavorative generali svolte, del raggruppamento “2-Professioni intellettuali, scientifiche e di elevata specializzazione” , “2.2 - Ingegneri, architetti e professioni assimilate”, .., fino al raggruppamento “2.2.1.6.1 - Ingegneri edili e ambientali”, producendo l’output della figura 19.

Figura 19 - Output della navigazione gerarchica della CP2011 per individuare la Professione di “Ingegnere edile”



Il sistema per la codifica delle professioni presente sul sito http://cp2011.istat.it ricorre ad una logica di ricerca, affine all’algoritmo di Google, basata sul potere discriminante delle stringhe e presenta un’aggregazione dei risultati che consente agli utenti di eseguire una scelta consapevole.

Nel dizionario dei nomi di Professioni all’interno del quale viene svolta la ricerca non sono state volutamente implementate le stringhe testuali che, seppur usate nel linguaggio comune, non hanno, nel linguaggio della classificazione, un significato univoco. Se, infatti, la ricchezza del dizionario delle Professioni può rappresentare un anello di congiunzione tra il linguaggio parlato dagli individui e quello utilizzato dalla classificazione, nel caso delle Professioni ciò crea un aumento considerevole dei falsi risultati positivi con i quali l’utente deve confrontarsi, a meno di non dettagliare appropriatamente le espressioni di linguaggio comune così da eliminarne l’eventuale ambiguità.

Tale rischio, infatti, è già insito in alcuni termini generalmente utilizzati nel linguaggio comune per descrivere la Professione.

Un esempio chiarificatore è rappresentato dal termine “meccanico”. Nel linguaggio comune, il “meccanico” viene generalmente associato al riparatore d’auto, ma all’interno della Classificazione delle Professioni questa stringa testuale può trovare numerose declinazioni, che vanno ad esempio dall’ingegnere meccanico, al perito meccanico, al disegnatore meccanico, al meccanico aeronautico e così via. Di conseguenza non essendo possibile abbinare la stringa ‘meccanico’ ad un codice univoco, sono state implementate nel dizionario tutte le stringhe che utilizzano il termine meccanico associandolo ad un successivo termine discriminante, in grado di ricondurre il nome ricercato ad un solo luogo della classificazione.

7.3 Il navigatore a supporto del quesito censuario

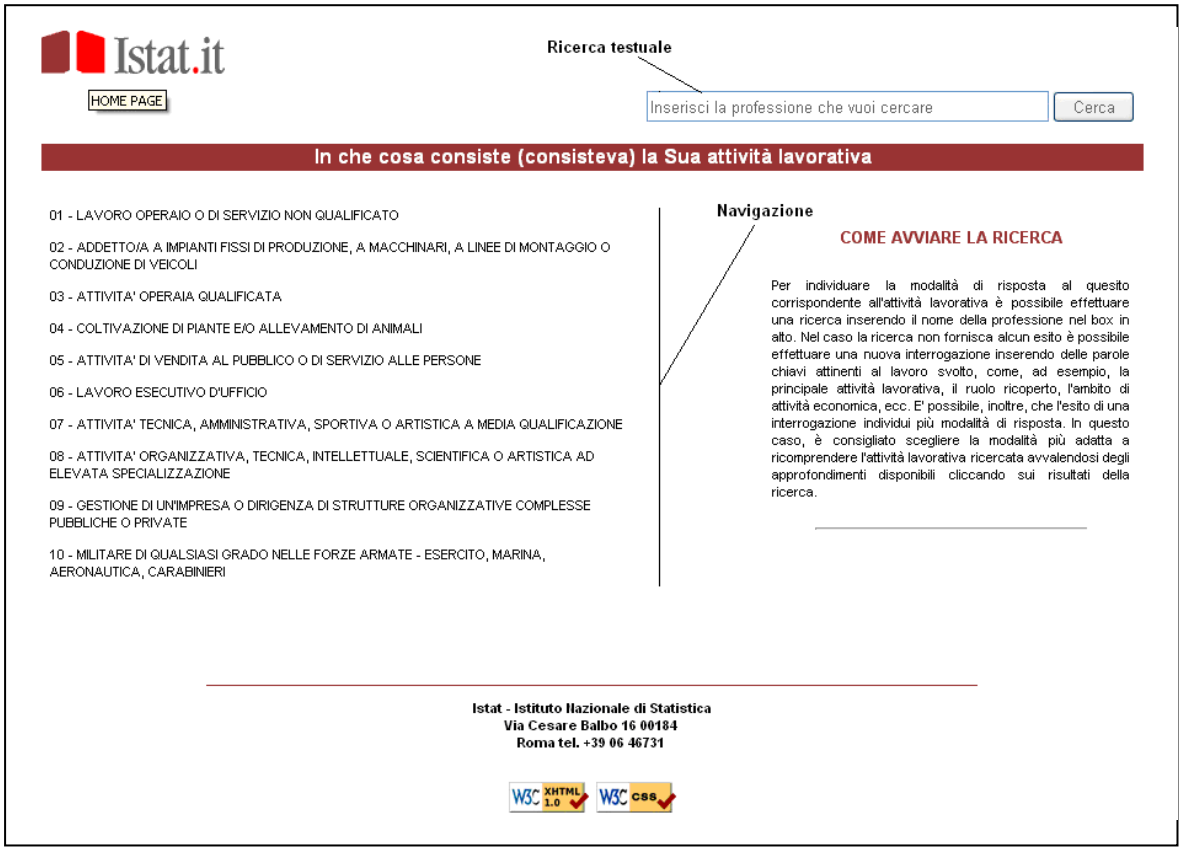
Per supportare l'utente nella compilazione del quesito censuario dedicato alla Professione non è stato possibile utilizzare direttamente il navigatore della Classificazione delle Professioni ma si è proceduto a sviluppare un apposito applicativo consultabile nel sito http://professioni.istat.it/censimenti/index.php.

Tale applicativo è stato realizzato nel linguaggio PHP5 e XHTML strict 1.0; il codice è stato validato con gli strumenti del W3C nel rispetto degli standard WCAG2.0 e dei requisiti della legge 4/2004 sull'accessibilità.

Poiché il quesito censuario richiedeva di convogliare tutte le Professioni svolte nel mercato del lavoro all'interno di 10 raggruppamenti, corrispondenti al primo livello classificatorio della Isco08, si è resa necessaria la predisposizione di una nuova base informativa.

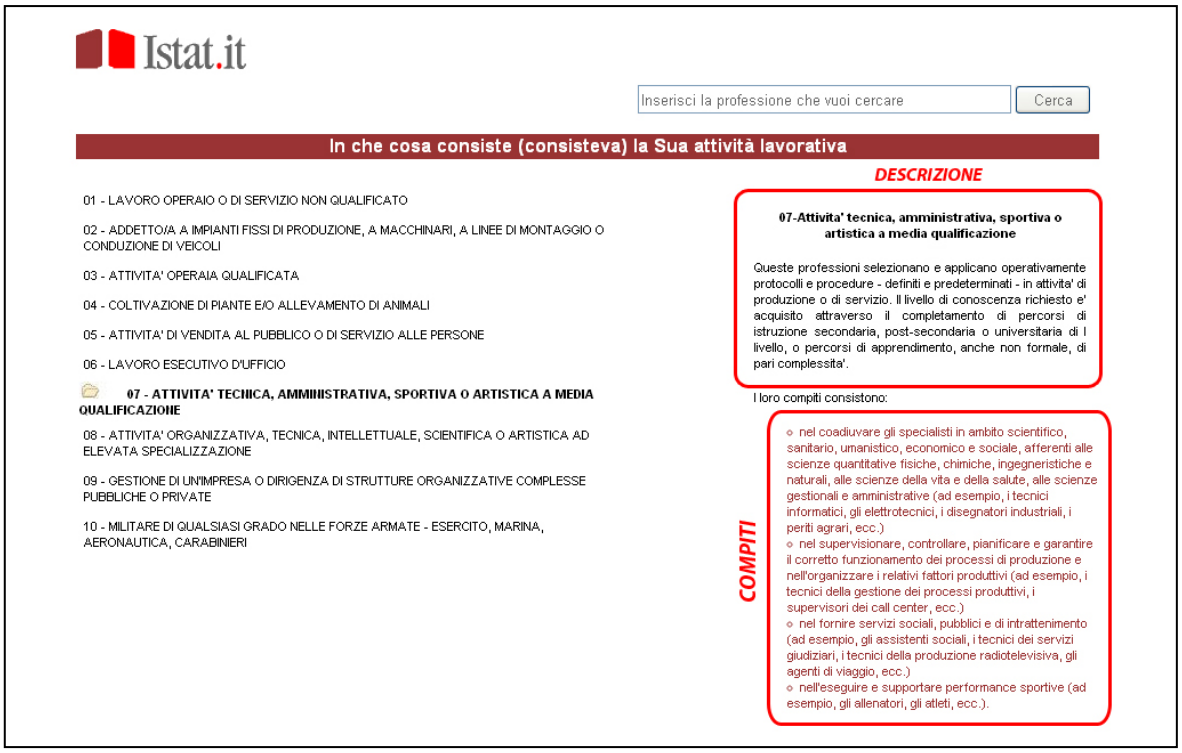
Come nel caso del navigatore della Classificazione CP2011, con il navigatore predisposto per la compilazione del quesito censuario l'utente aveva la possibilità di navigare tra le opzioni proposte per approfondirne il significato oppure di effettuare una ricerca libera di un campo testuale (Figura 20). È stata predisposta un'interfaccia di consultazione intuitiva che ha consentito la scelta immediata di uno dei due percorsi.

Figura 20 - Home page del navigatore per il quesito censuario



Dalla *home page*, infatti, facendo clic su una modalità proposta dall'elenco, veniva presentata una descrizione dettagliata e un elenco dei compiti generalmente svolti dalle professioni racchiuse in quel particolare raggruppamento (Figura 21).

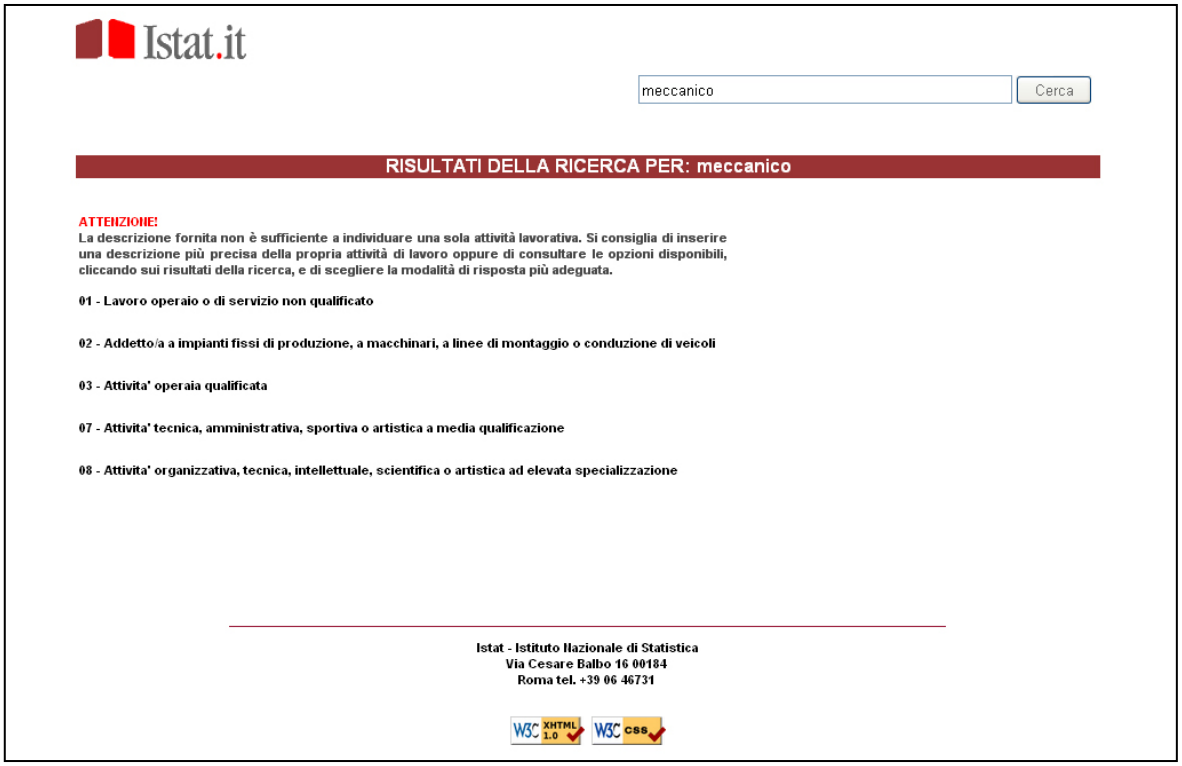
Figura 21 - Output della navigazione tra le opzioni proposte dal quesito censuario



La ricerca libera si basa su un algoritmo ad hoc che elabora le voci professionali, le ordina e le raggruppa all'interno dei raggruppamenti di appartenenza.

Così ad esempio la ricerca libera della stringa testuale "meccanico" fornisce l'*output* rappresentato nella figura 22.

Figura 22 – Output della ricerca della stringa testuale " meccanico"

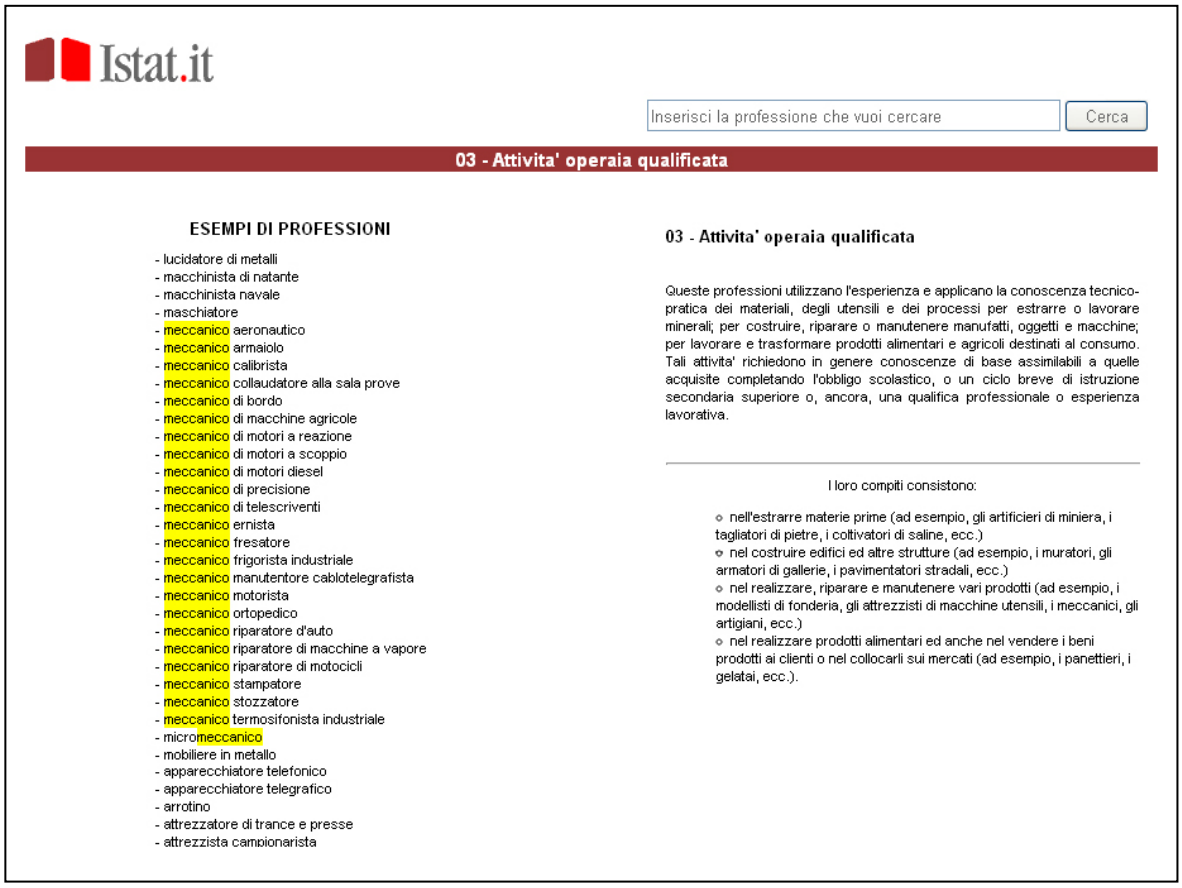


I risultati della ricerca sono raggruppati in base all'area tematica a cui appartengono. L'utente era pertanto obbligato a riflettere per scegliere l'area in cui collocare la stringa cercata.

Se stava cercando il "meccanico riparatore d'auto" escludeva certamente le professioni intellettuali o il lavoro operaio non qualificato e sceglieva l'opzione 03 – Attività operaia qualificata.

Facendo clic su questo gruppo, poteva consultare una scheda descrittiva nella quale sono elencate le voci professionali con la stringa cercata evidenziata, la descrizione e l'elenco dei compiti (Figura 23).

Figura 23 - Output della ricerca della stringa "meccanico" e della successiva scelta dell'opzione "03- Attività operaia qualificata"

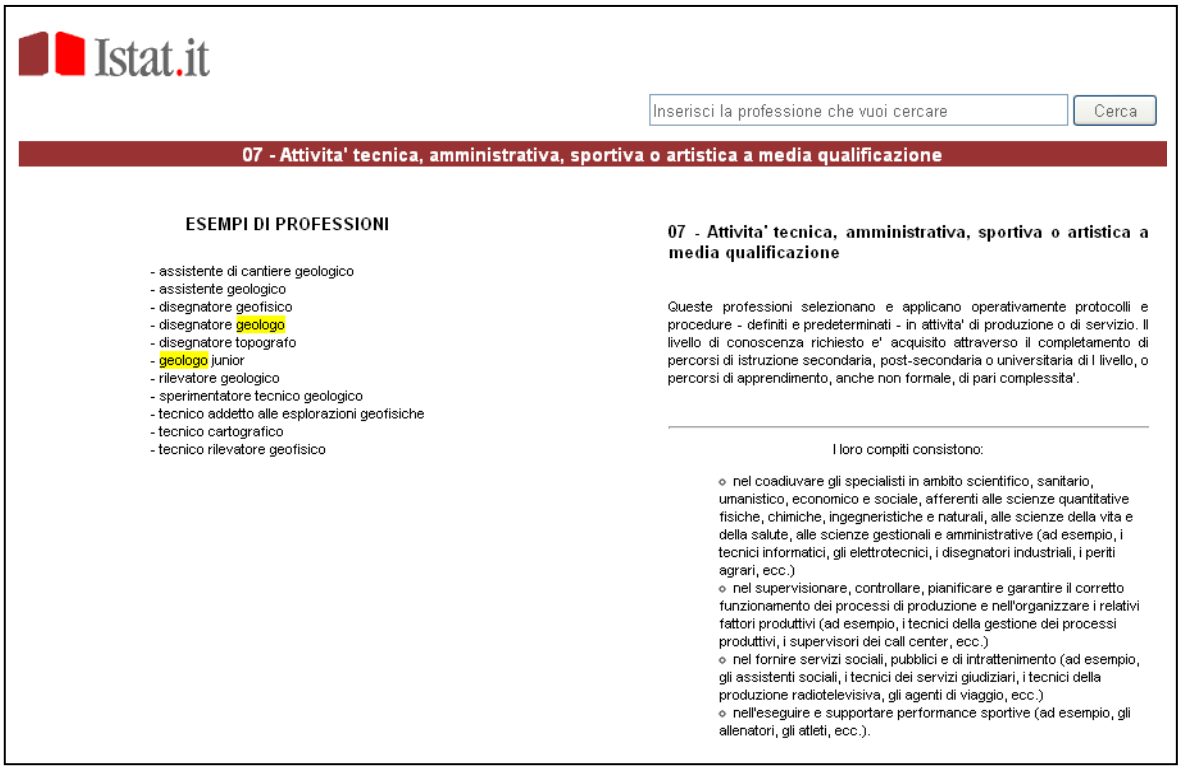


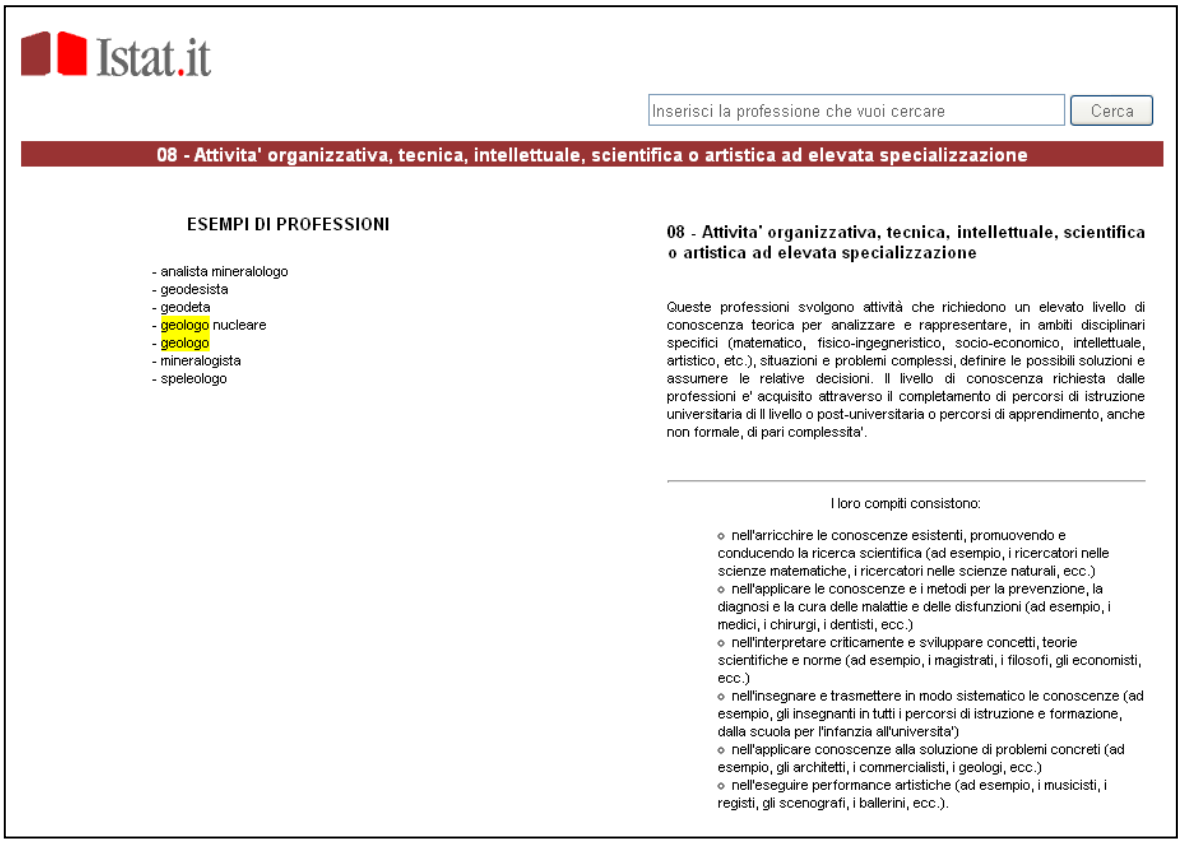
L'ausilio fornito ai rispondenti dal Navigatore di ricerca si evince dai diversi casi in cui i cittadini potevano rimanere disorientati nella scelta tra le possibili codifiche di una stessa Professione in base al diverso livello di competenza richiamato.

Si pensi, a titolo di esempio, al duplice sbocco di codifica previsto dalla Professione del 'geologo', che può essere esercitata in qualità di 'Geologo junior' – professione appartenente al raggruppamento delle attività tecniche a media qualificazione (07) - o di 'Geologo' – professione appartenente al raggruppamento delle attività tecniche a elevata specializzazione (08) -.

In questi casi, il Navigatore presentava entrambe le scelte di codifica al rispondente, illustrando le differenze tra i due esiti in termini di definizione dei due raggruppamenti professionali, di compiti previsti e di esempi di professioni ricomprese al loro interno, fornendo in altre parole tutte le informazioni essenziali per desumere il codice adatto all'occupazione in esame (Figura 24).

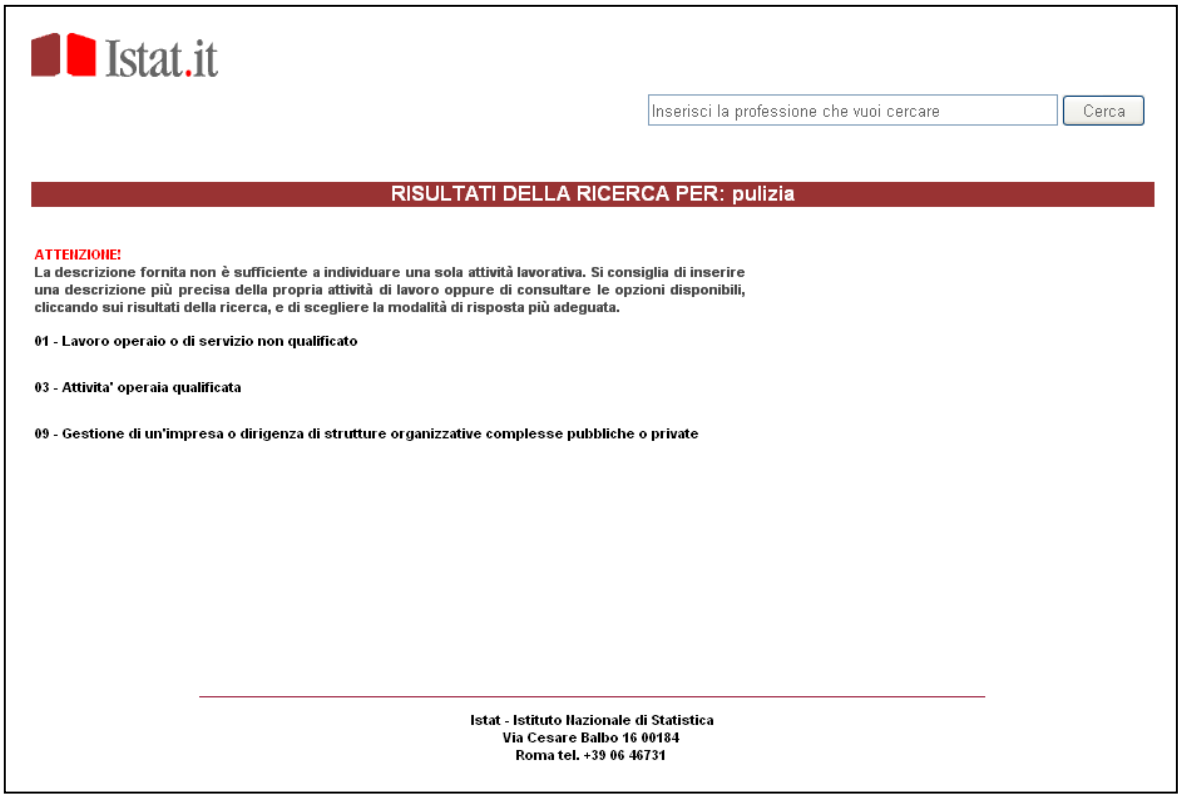
Figura 24 - Descrizione dei due possibili esiti di codifica per la professione di "geologo"





L'impiego del navigatore è risultato altrettanto informativo nei casi in cui i cittadini intendessero esplorare i possibili esiti di codifica delle professioni di un determinato ambito lavorativo. Si pensi, sempre a titolo di esempio, alla ricerca dei codici previsti nel settore delle pulizie. In questo caso il Navigatore presentava una molteplicità di esiti legati sia ai diversi ruoli che è possibile svolgere nel settore (imprenditore piuttosto che addetto alle pulizie), sia ai diversi livelli di qualificazione previsti per eseguire certe mansioni (attività di pulizia qualificata vs. non qualificata) (Figura 25).

Figura 25 - Esito della ricerca della stringa testuale "pulizia"



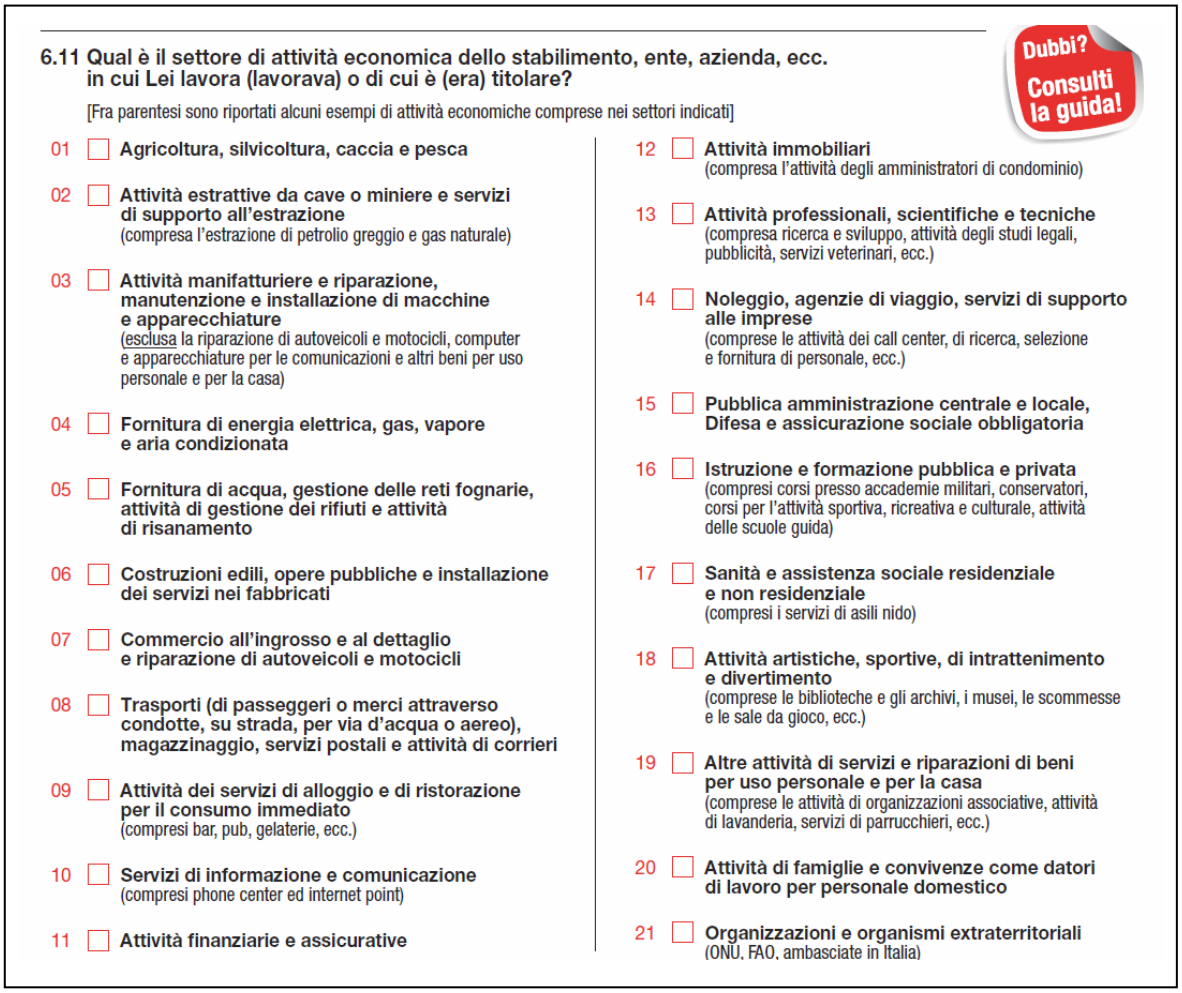
8. Il Navigatore delle Attività economiche

8.1 La base informativa del navigatore ATECO

Come già sottolineato nel paragrafo 2, analogamente al Censimento del 2001, nel questionario del 2011 non è stato previsto di rilevare l'informazione dell'Attività economica espletata dal rispondente con una risposta a testo libero; i costi di acquisizione e trattamento di questa tipologia di risposta, nonché il non elevato livello di qualità ottenibile (risposte spesso generiche o non appropriate) hanno portato a decidere di acquisire questa informazione con un quesito precodificato che consentisse l'individuazione del settore Ateco, senza scendere ad un ulteriore livello di dettaglio.

In particolare, il quesito, come nella figura di seguito riportata, prevede ventuno modalità corrispondenti alle 21 Sezioni della Classificazione internazionale delle Attività economiche NACE Rev.2 (recepita in Italia dalla Ateco 2007), in accordo con quanto richiesto dalla normativa vigente dell'Unione Europea in materia di Censimenti.[17]

Figura 26 - Quesito per rilevare l'Attività economica nel modello censuario 2011



6.11 Qual è il settore di attività economica dello stabilimento, ente, azienda, ecc. in cui Lei lavora (lavorava) o di cui è (era) titolare?

Dubbi? Consulti la guida!

[Fra parentesi sono riportati alcuni esempi di attività economiche comprese nei settori indicati]

01 ☐ **Agricoltura, silvicoltura, caccia e pesca**

02 ☐ **Attività estrattive da cave o miniere e servizi di supporto all'estrazione**
(compresa l'estrazione di petrolio greggio e gas naturale)

03 ☐ **Attività manifatturiere e riparazione, manutenzione e installazione di macchine e apparecchiature**
(esclusa la riparazione di autoveicoli e motocicli, computer e apparecchiature per le comunicazioni e altri beni per uso personale e per la casa)

04 ☐ **Fornitura di energia elettrica, gas, vapore e aria condizionata**

05 ☐ **Fornitura di acqua, gestione delle reti fognarie, attività di gestione dei rifiuti e attività di risanamento**

06 ☐ **Costruzioni edili, opere pubbliche e installazione dei servizi nei fabbricati**

07 ☐ **Commercio all'ingrosso e al dettaglio e riparazione di autoveicoli e motocicli**

08 ☐ **Trasporti (di passeggeri o merci attraverso condotte, su strada, per via d'acqua o aereo), magazzinaggio, servizi postali e attività di corrieri**

09 ☐ **Attività dei servizi di alloggio e di ristorazione per il consumo immediato**
(compresi bar, pub, gelaterie, ecc.)

10 ☐ **Servizi di informazione e comunicazione**
(compresi phone center ed internet point)

11 ☐ **Attività finanziarie e assicurative**

12 ☐ **Attività immobiliari**
(compresa l'attività degli amministratori di condominio)

13 ☐ **Attività professionali, scientifiche e tecniche**
(compresa ricerca e sviluppo, attività degli studi legali, pubblicità, servizi veterinari, ecc.)

14 ☐ **Noleggio, agenzie di viaggio, servizi di supporto alle imprese**
(comprese le attività dei call center, di ricerca, selezione e fornitura di personale, ecc.)

15 ☐ **Pubblica amministrazione centrale e locale, Difesa e assicurazione sociale obbligatoria**

16 ☐ **Istruzione e formazione pubblica e privata**
(compresi corsi presso accademie militari, conservatori, corsi per l'attività sportiva, ricreativa e culturale, attività delle scuole guida)

17 ☐ **Sanità e assistenza sociale residenziale e non residenziale**
(compresi i servizi di asili nido)

18 ☐ **Attività artistiche, sportive, di intrattenimento e divertimento**
(comprese le biblioteche e gli archivi, i musei, le scommesse e le sale da gioco, ecc.)

19 ☐ **Altre attività di servizi e riparazioni di beni per uso personale e per la casa**
(comprese le attività di organizzazioni associative, attività di lavanderia, servizi di parrucchieri, ecc.)

20 ☐ **Attività di famiglie e convivenze come datori di lavoro per personale domestico**

21 ☐ **Organizzazioni e organismi extraterritoriali**
(ONU, FAO, ambasciate in Italia)

[17] COMMISSION REGULATION (EC) No 1201/2009 of 30 November 2009 implementing Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses as regards the technical specifications of the topics and of their breakdowns.

Con tale soluzione, pur semplificando la gestione di questa variabile rispetto alla rilevazione con un quesito aperto, permaneva tuttavia una possibile difficoltà da parte del rispondente nel collocarsi in una delle ventuno modalità che, essendo necessariamente descritte in modo molto sintetico, non potevano esplicitare casistiche di dettaglio spesso corrispondenti alla peculiare attività confacente al rispondente.

Per fare un esempio banale, non è immediatamente deducibile che *l'attività giornalistica* rientri nel ramo *18 - Attività artistiche, sportive, di intrattenimento e divertimento*, oppure che *l'attività di cura e custodia di bambini (baby sitter)* rientri nel ramo *17 - Sanità e assistenza residenziale e non residenziale*.

Al fine quindi di facilitare l'individuazione del settore Ateco di propria competenza, è stato deciso di mettere in linea un navigatore nella classificazione che consentisse di individuare la modalità di risposta a partire da una descrizione espressa con il linguaggio naturale abitualmente utilizzato dagli utenti.

Poiché l'Istat già disponeva di un navigatore con queste caratteristiche, in linea sul sito istituzionale dell'Istituto, si è deciso di avvalersi di questo strumento, adattandolo alle necessità del censimento della Popolazione.

Il navigatore delle Attività economiche di cui si è parlato è accessibile sul sito www.istat.it da giugno 2008 (2009, Istat) e registra una media di oltre 15.000 accessi a settimana (2012, Ferrillo et al.), in quanto sono molteplici gli utenti interessati ad individuare un codice Ateco a partire da una descrizione libera dell'attività espletata, tanto più che la nuova classificazione Ateco del 2007 è l'unica utilizzata da tutti gli Enti interessati (Agenzia delle Entrate, Camere di Commercio, Inail e Inps). Per realizzare questa funzione ci si è avvalsi dell'algoritmo di *matching* implementato nel *software* di codifica automatica ACTR, nonché della base informativa Ateco messa a punto per codificare i dati rilevati in molte indagini Istat. Per dare un ordine di grandezza in merito a quanto questa base informativa sia stata arricchita per avvicinarsi il più possibile al modo di esprimersi dei rispondenti, si evidenzia che, a fronte di 1893 descrizioni della classificazione ufficiale, il dizionario ne comprende oltre 33.000. Obiettivo primario di questa funzione è quello di individuare un codice Ateco al massimo dettaglio, vista la tipologia di utente cui è rivolta; nel caso del censimento della popolazione, invece, avere un tale livello di dettaglio sarebbe stato probabilmente confondente per il rispondente cui era richiesto soltanto di collocarsi in una delle ventuno sezioni della classificazione.

L'adattamento del navigatore disponibile sul sito istituzionale quindi, oltre ad essere stato funzionale ad esigenze di tipo architetturale e sistemistico (si veda paragrafo 8.2), ha risposto a questa necessità di ricondurre le risposte testuali ad un così elevato livello di aggregazione.

Si riportano di seguito nelle figure 27 e 28 le risposte delle due versioni di navigatore alla stessa *query = giornalista.* Come può vedersi, con la funzione di navigazione nella Classificazione Ateco 2007 accessibile dal sito www.istat.it si ottengono tre codici al massimo dettaglio tra i quali l'utente *Web* può individuare quello confacente alla propria attività (90.03.01, 90.03.02 e 90.03.09); tutti e tre questi codici corrispondono allo stesso ramo Ateco (18) di cui viene effettuato il *display* con il navigatore messo a punto per il censimento della popolazione.

Si specifica infine che tale funzione è stata resa linkabile dal Sito del Censimento, selezionando il box Strumenti, ma anche direttamente dalla guida alla compilazione del questionario, cliccando su un apposito box in corrispondenza delle spiegazioni di dettaglio sul quesito 6.11.

Figura 27 - Risposta alla *query* "giornalista" del navigatore su www.istat.it

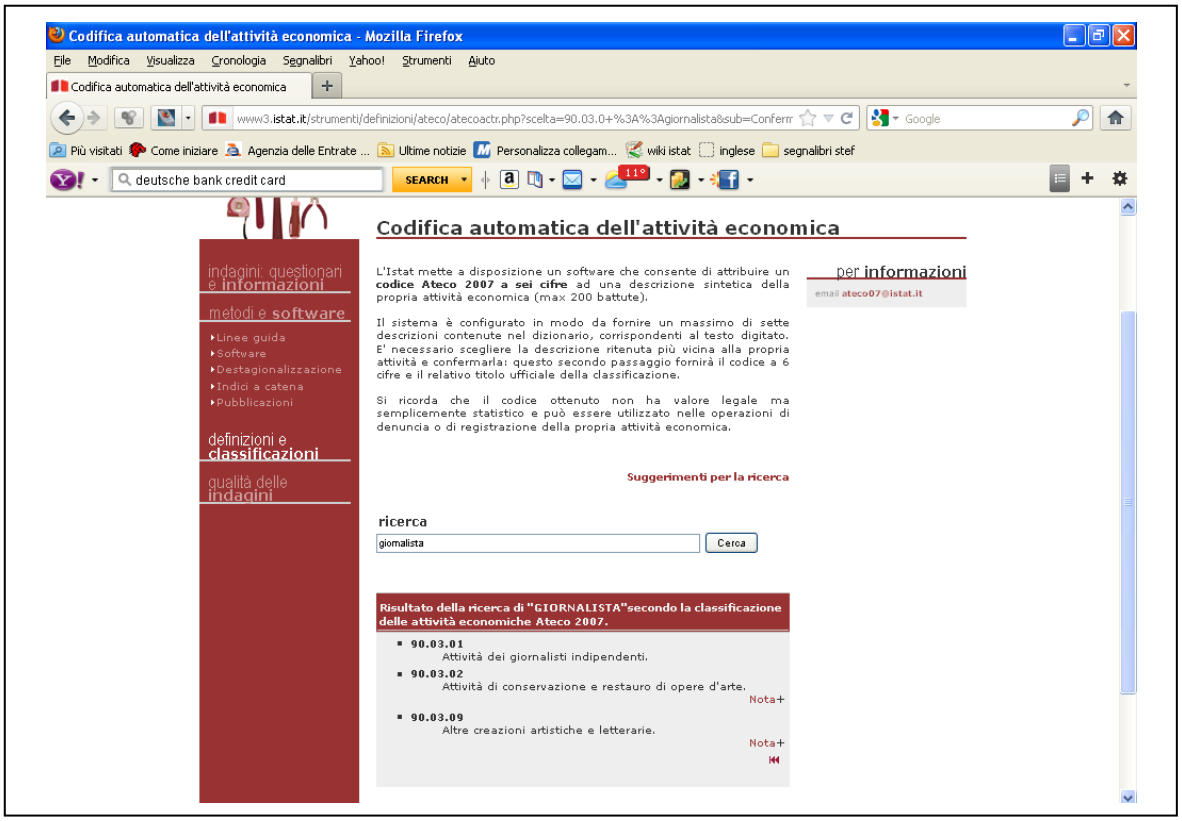
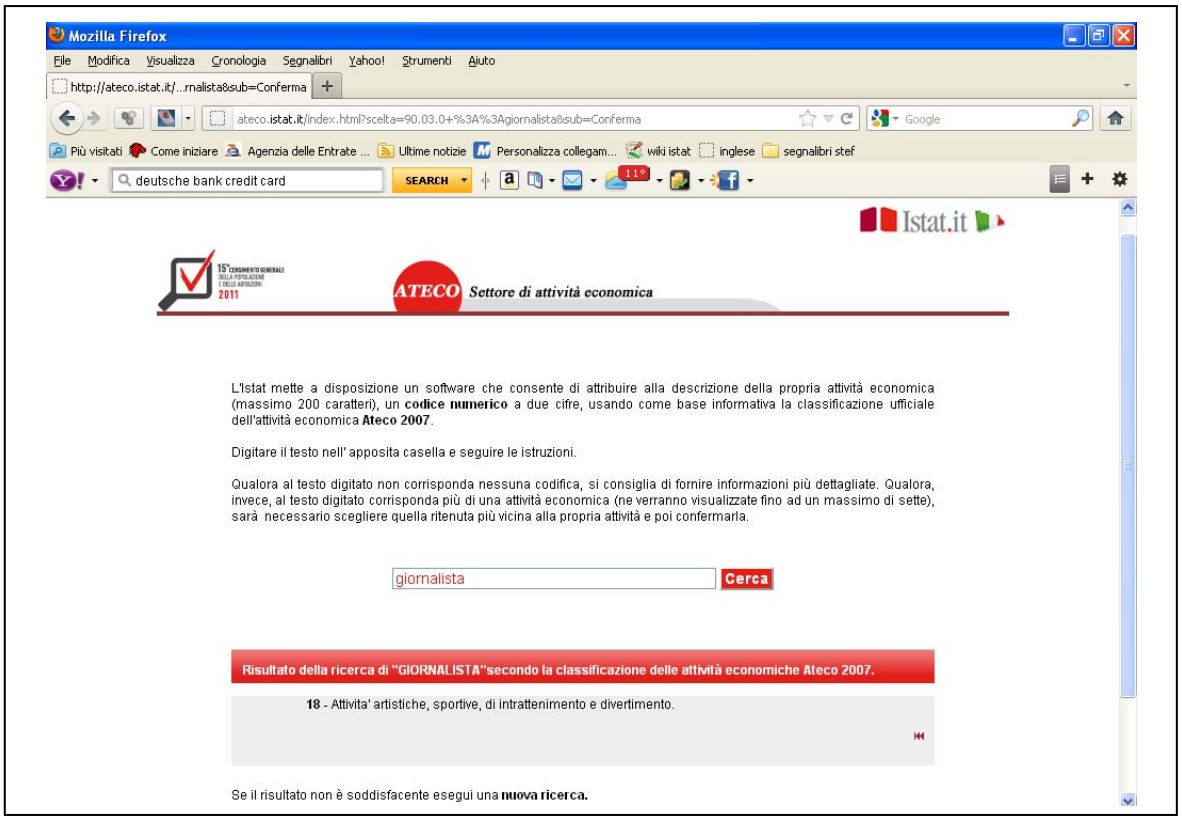


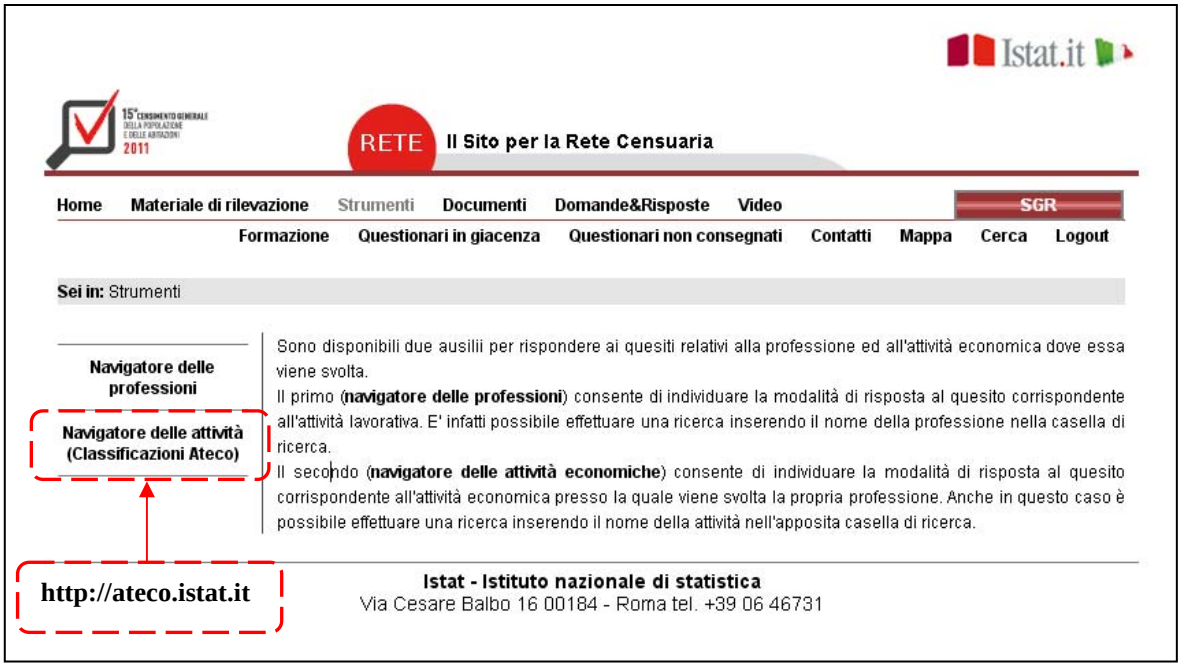
Figura 28 - Risposta alla *query* "giornalista" del navigatore messo a punto per il Censimento della popolazione



8.2 Il navigatore per il Censimento

L'utente poteva accedere al navigatore sia attraverso il Sito della Rete Censuaria (in caso di operatore abilitato) (Figura 29) sia attraverso il sito dell'Istituto (www.istat.it).

Figura 29 - Il Sito della Rete Censuaria

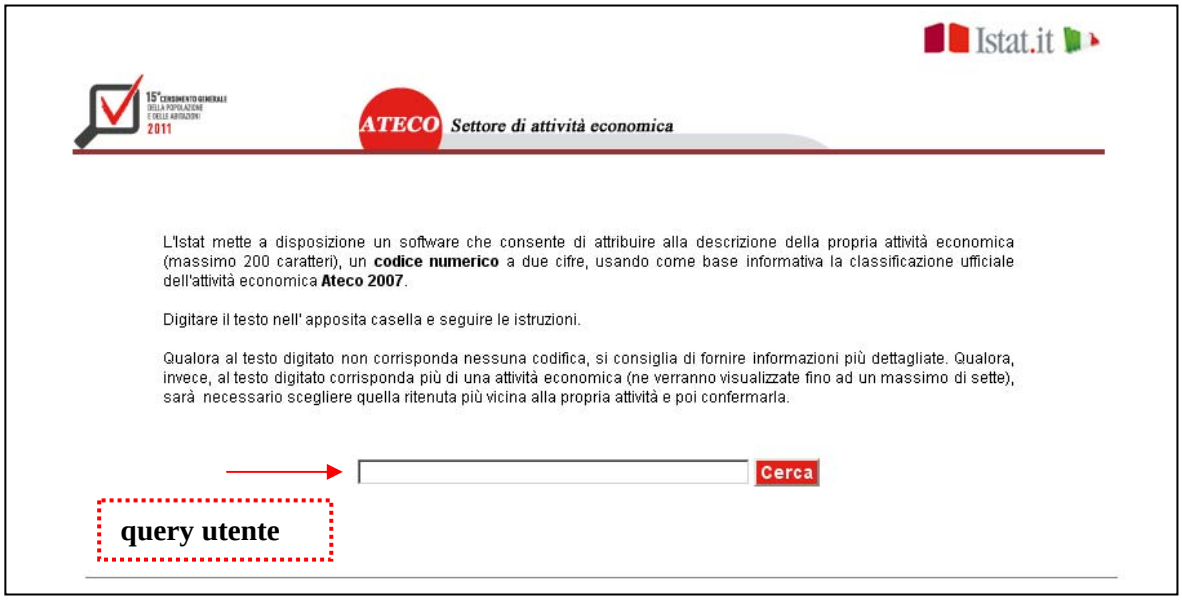


Il sito **http://ateco.istat.it** è costituito da pagine HTML dinamiche scritte in PHP e consta essenzialmente di due parti.

La **prima parte** è costituita da pagine Web, che risiedono sui server OHIO e IOWA (S.O. Linux); in particolare sono:

- la pagina di presentazione dello strumento (Figura 30) in cui venivano acquisite le *query* dell'utente;
- la pagina di pubblicazione dei risultati ottenuti con il sistema ACTR. Questo sistema è configurato in modo da fornire in *output* un elenco di descrizioni delle possibili attività (da 1 a 7 *item*); l'utente doveva poi scegliere da quest'elenco la descrizione che riteneva più vicina alla propria Attività economica e selezionare il bottone di conferma (Figura 31). Nel caso in cui alla *query* dell'utente non corrispondesse nessuna Attività economica, il sistema ACTR generava un messaggio di ricerca fallita (Figura 33);
- la pagina di pubblicazione dei risultati finali che forniva il codice a 2 cifre e il relativo titolo ufficiale della classificazione (Figura 32).

Figura 30 - Pagina iniziale dell'applicazione



La *query* utente è la descrizione sintetica (massimo 200 caratteri) dell'Attività economica che veniva successivamente elaborata dal software ACTR per la codifica automatica del testo.

Figura 31 - Risultato della *query* utente

La funzione di "*nuova ricerca*" consentiva di ricaricare la pagina iniziale eliminando tutti i dati digitati precedentemente dall'utente per avviare quindi una nuova sessione di ricerca.

Figura 32 - Output finale della ricerca: codifica ufficiale delle Attività economiche Ateco 2007

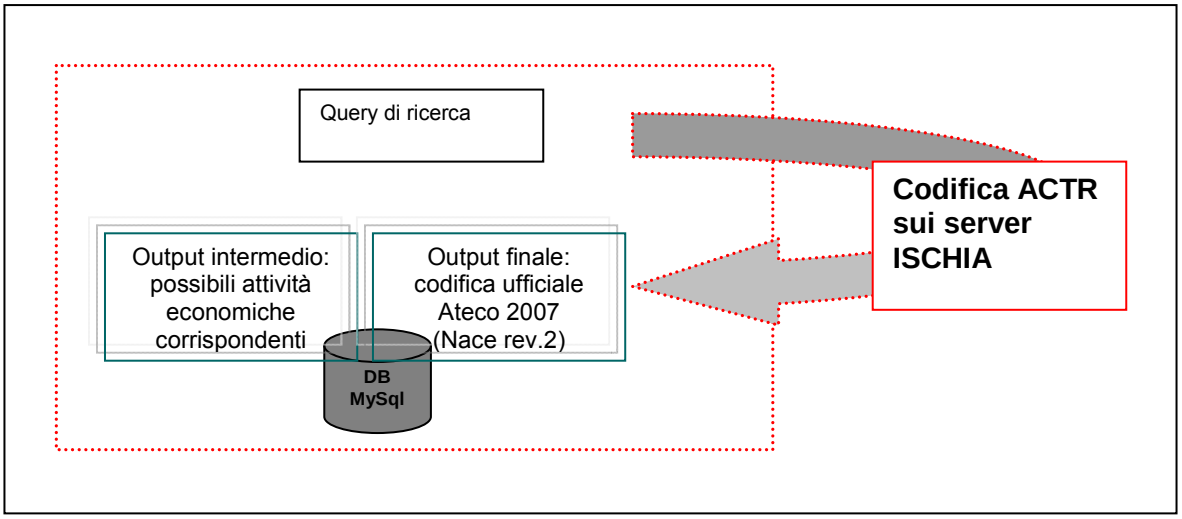
Il tasto di "navigazione" ⏮ permetteva di passare da quest'ultima pagina di *output* a quella intermedia (Figura 31) per dare la possibilità all'utente di fare una diversa selezione dall'elenco delle classificazioni proposte da ACTR.

Figura 33 - Ricerca fallita

Se la stringa di ricerca digitata dall'utente era errata, il sistema visualizzava la pagina mostrata in Figura 33 contenente le indicazioni per effettuare una corretta ricerca.

La **seconda parte** dell'applicazione comprende i programmi in PHP che si trovano sul server ISCHIA su cui gira il software ACTR. Questi programmi, chiamati in *http* dalle pagine del sito, avevano le funzioni di costruire gli *input* per il software ACTR, eseguire il software, recuperare gli *output*, formattarli e farli pubblicare nella pagina mostrata nella Figura 31. Si noti che il *software* ACTR, come descritto nel successivo paragrafo, è tale che non può accettare più richieste in parallelo, per cui è stato necessario serializzare le richieste utente provenienti dalla pagina *Web* principale.

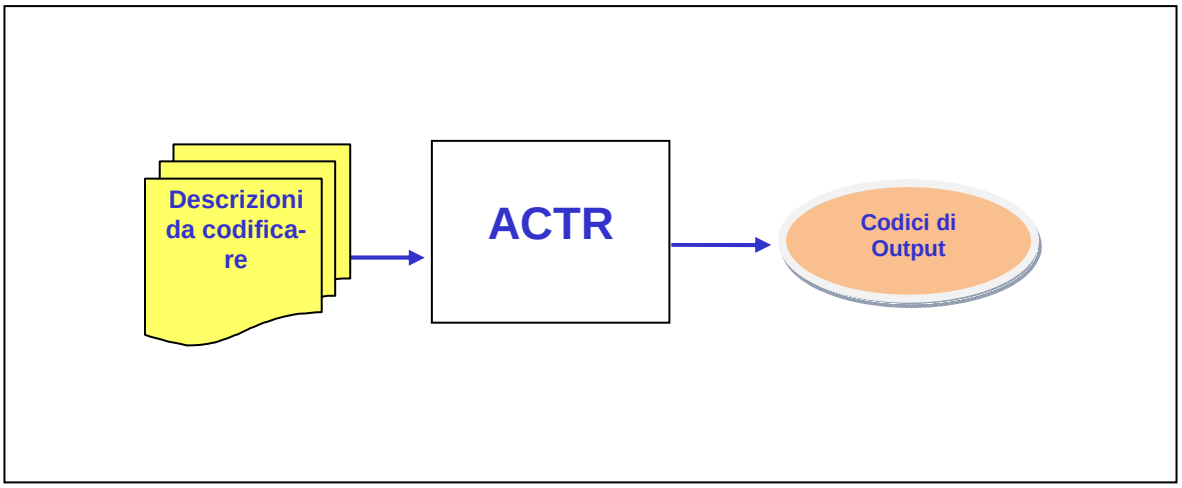
Il flusso dell'intera applicazione è schematizzato come segue:



Nell'ultima fase, l'applicazione prendeva il codice ACTR (di cinque *digit*) selezionato dall'utente e lo confrontava con i dati memorizzati sulla tabella "***user_ateco_actr"*** del corrispondente DB MySQL. Il risultato di questa *select* era visualizzato in output (Figura 31 – Risultato della *query* utente) dove venivano forniti all'utente i codici e le corrispondenti descrizioni delle attività.

8.2.1 Integrazione del software ACTR nell'applicazione Web

L'applicativo ACTR legge il *file* di *input* che viene creato dinamicamente dall'applicazione *PHP* con le *query* utente, elabora questo *file* utilizzando opportune regole di *parsing*, trova il/i codice/i e la relativa descrizione che corrispondono alla *query* digitata dall'utente. I risultati vengono scritti alternativamente in uno dei quattro *file* ASCII, a seconda del tipo di corrispondenza trovata. L'applicazione *Web* legge i *file* forniti da ACTR, li decodifica e visualizza sulla pagina con un numero massimo di 7 righe nel caso di risultati multipli o possibili. Nel caso invece ACTR non trovi nessuna associazione viene fornito un opportuno messaggio (Figura 33).

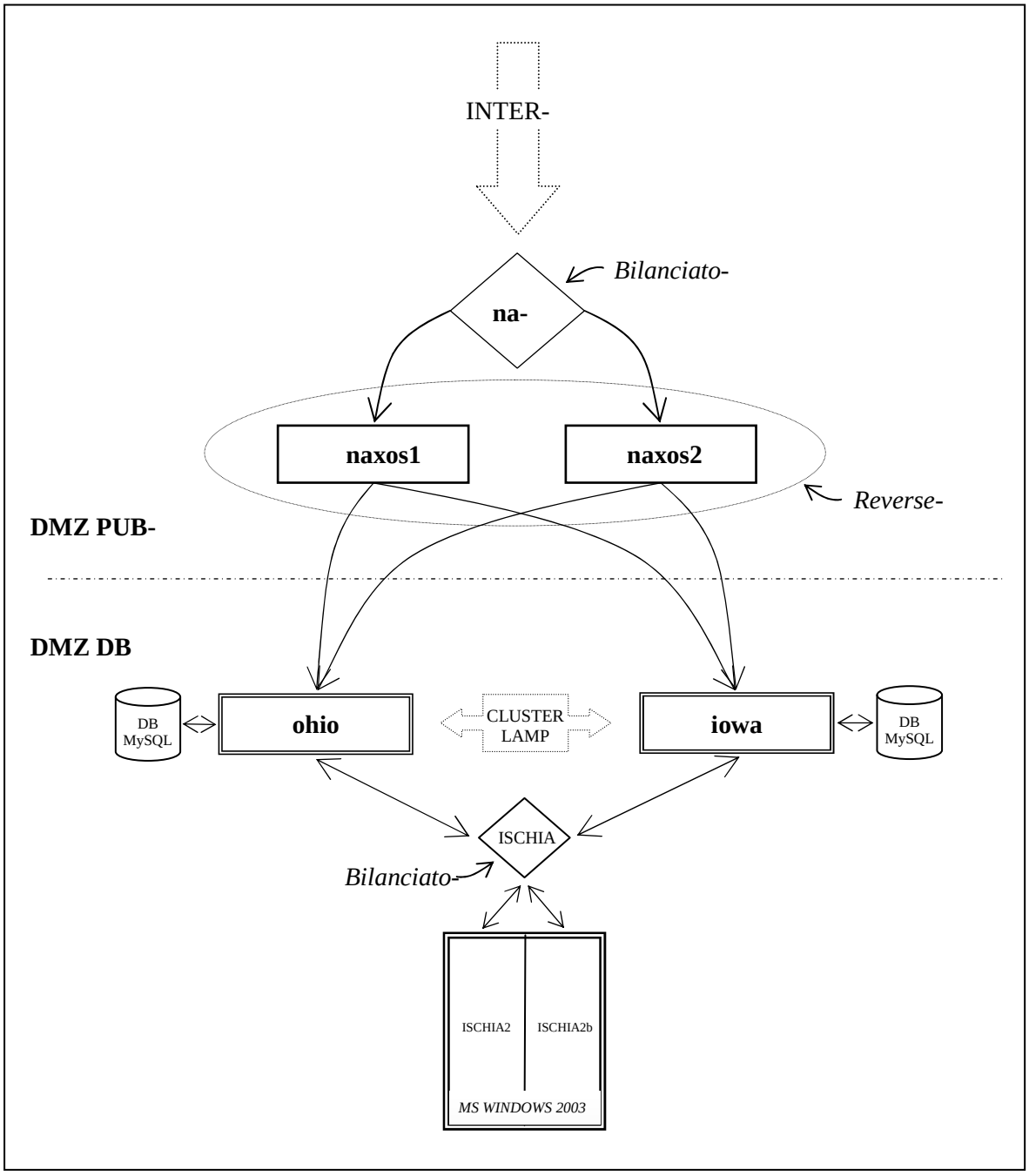


8.3 L'architettura del sito http://ateco.istat.it

L'indirizzo Web http://ateco.istat.it puntava (tramite DNS) al bilanciatore di carico naxos, che distribuiva le richieste http verso i reverse-proxy naxos1 e naxos2, i quali inoltravano il traffico http sui server ohio e iowa (dmz-db). Ciascuno di questi due server ospitava un Web-server apache+php e un DB MySQL.

Lo schema che segue mostra l'architettura complessiva del sito.

Figura 34 - Architettura del sito http://ateco.istat.it



I due reverse-proxy naxos1 e naxos2 inoltravano le richieste http verso gli application-server ohio e iowa, bilanciando il traffico mediante il modulo mod_proxy_balancer di apache 2.2. Il meccanismo consente inoltre la persistenza della connessione client in modo tale da non invalidarne la sessione.

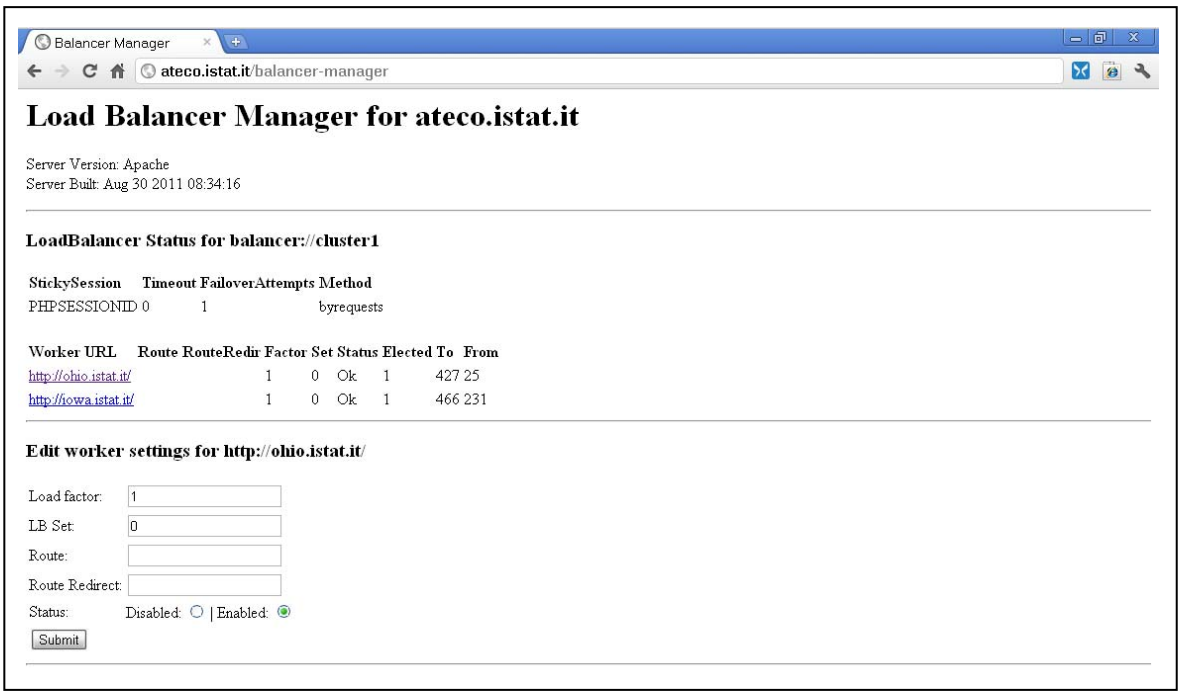
Gli scopi di questo meccanismo sono:

- mantenere la disponibilità del sito *Web* (in caso di problemi su uno dei server che ospitano il sito, il traffico viene raccolto dagli altri server);
- consentire il bilanciamento del carico (il flusso di rete e il carico di sistema nel back-end viene distribuito tra le macchine che compongono il *cluster*);
- fornire una replica di sicurezza (i due server sono costantemente allineati fornendo quindi un'ulteriore copia di backup).

Il mod_proxy_balancer offre anche la possibilità di utilizzare una interfaccia *Web* di gestione, configurabile all'interno dello stesso virtualhost del sito. Nel caso del sito Ateco, l'accesso al servizio era riservato alla rete interna e protetto con *password*. Questo perché l'interfaccia non era soltanto informativa, ma poteva essere utilizzata per distribuire, regolare o interrompere il traffico http verso gli application server.

La pagina *Web* di gestione del sito ateco.istat.it appare come quella dell'immagine seguente (Figura 35), in cui risulta evidente la possibilità di modificare alcuni parametri come il fattore di carico o lo status.

Figura 35 - Load Balancer Manager per il sito http://ateco.istat.it



8.4 Test di carico dell'applicazione http://ateco.istat.it

La tavola 27 che segue mostra i risultati ottenuti durante i test di performance dell'applicazione ACTR realizzata per il sito http://ateco.istat.it , ottenuti utilizzando il prodotto JMeter.

I test sono stati effettuati simulando un numero da 10 fino a 300 richieste (processi) contemporanei e soggetti al traffico esistente sulla rete durante il periodo di simulazione.

I dati riportati nella tavola sono:

- il tempo massimo, minimo e medio di risposta di un campione;
- la deviazione standard, che rappresenta la misura di quanto si disperdono i valori di un campione rispetto al valore medio (una deviazione standard elevata indica grandi variazioni nei tempi di risposta);
- il throughput, che esprime la quantità di dati trasmessi in una unità di tempo, in questo caso 1 secondo;
- la % di errore nella rilevazione dei dati.

I tempi di risposta e la deviazione standard sono espressi in millisecondi.

Tavola 27 - Tempi di risposta per numero di accessi contemporanei

NUMERO DI PROCESSI CONTEMPORANEI	10	50	100	300
Tempo massimo di risposta	8080	67985	72041	186150
Tempo minimo di risposta	191	191	208	195
Tempo medio di risposta	1590	6018	10044	24787
Deviazione standard	1100	7826	12751	33789
Throughput	5.7/sec	6.2/sec	5.7/sec	7/sec
% di errore	0.00%	0.00%	0.00%	0.00%

Da una prima analisi si evince che i tempi di risposta, passando da un numero di richieste contemporanee di 10 ad un numero di 300 varia nel seguente modo:

Numero di processi contemporanei	10	>>>	300	
Tempo massimo di risposta	8 secondi	>>>	3	minuti
Tempo minimo di risposta	0.19 secondi	>>>	0.2	secondi
Tempo medio di risposta	1.5 secondi	>>>	25	secondi

Si fa notare che con richieste contemporanee si intende che 10, 100…1000 utenti contemporaneamente cliccano sul bottone *Cerca* dell'applicazione *Web*.

I grafici che seguono (Figure da 36 a 43) rappresentano rispettivamente:

- istogramma di distribuzione dei tempi di risposta (in ms);
- grafico della distribuzione in percentuale (percentili) dei tempi di risposta (in ms).

Figura 36 - Distribuzione dei tempi di risposta per 10 richieste contemporanee

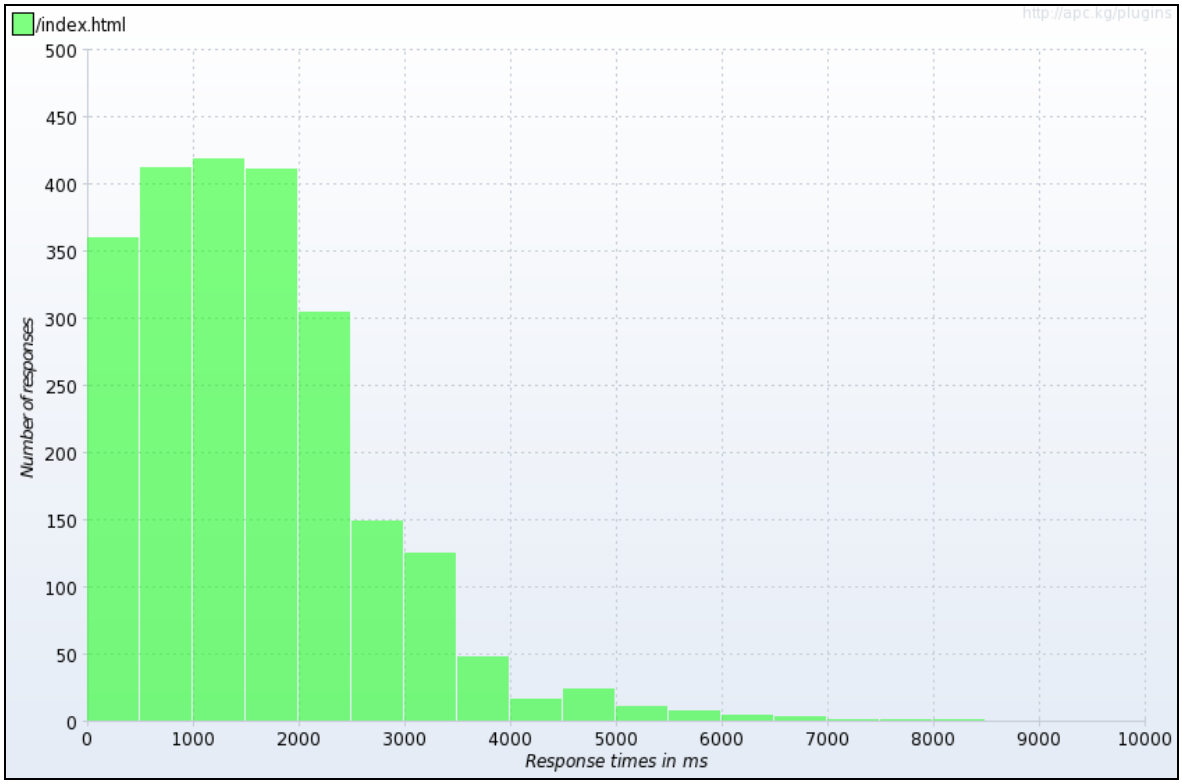


Figura 37 - Tempi di risposta in percentili per 10 richieste contemporanee

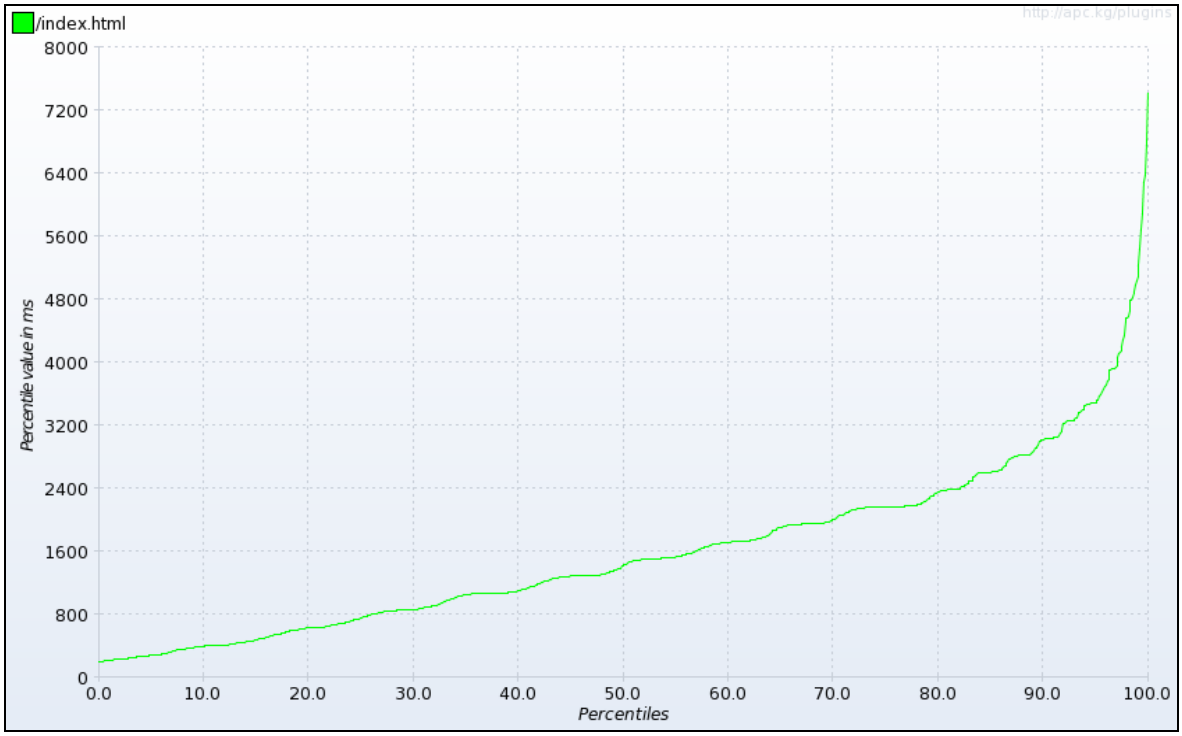


Figura 38 - Distribuzione dei tempi di risposta per 50 richieste contemporanee

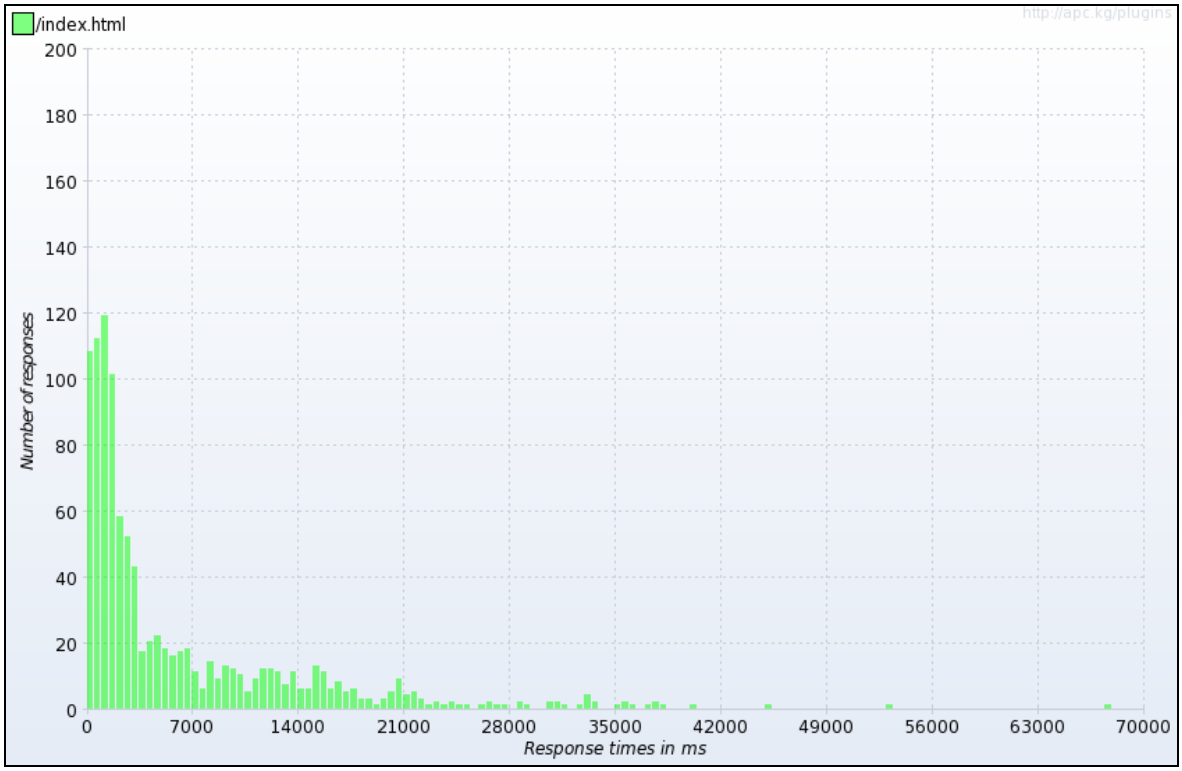


Figura 39 - Tempi di risposta in percentili per 50 richieste contemporanee

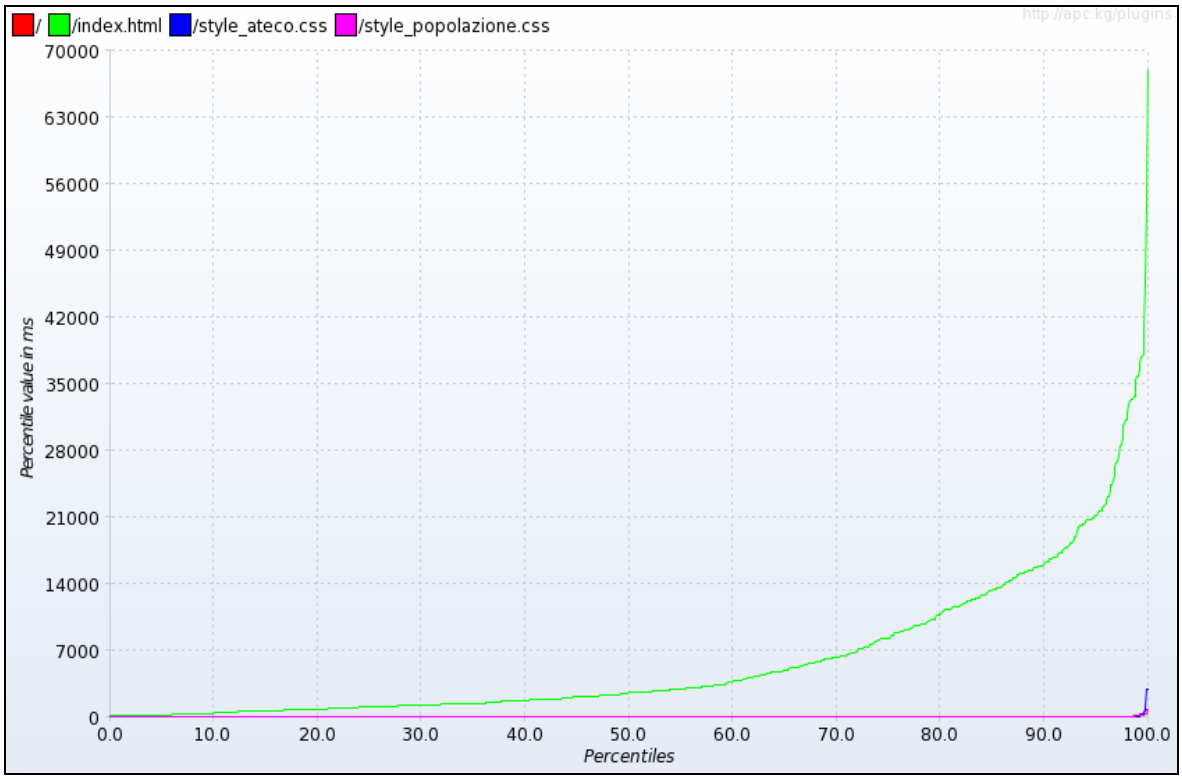


Figura 40 - Distribuzione dei tempi di risposta per 100 richieste contemporanee

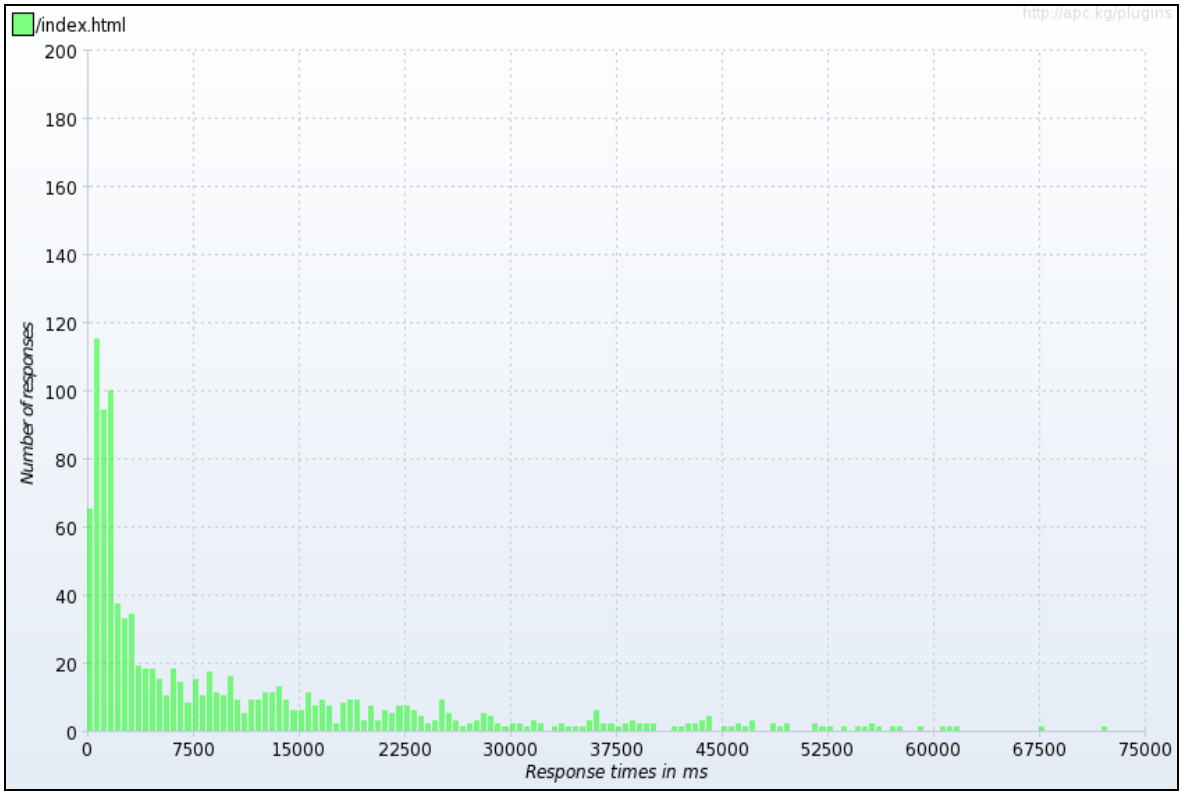


Figura 41 - Tempi di risposta in percentili per 100 richieste contemporanee

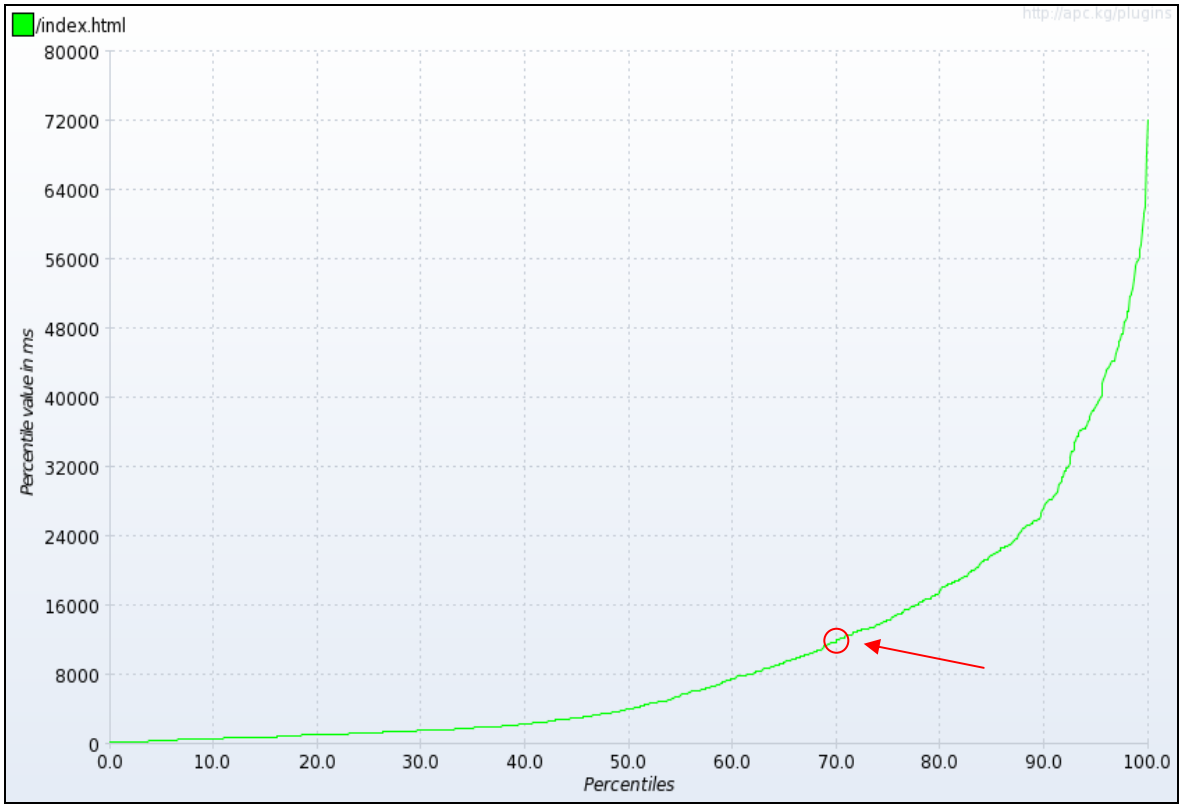


Figura 42 - Distribuzione dei tempi di risposta per 300 richieste contemporanee

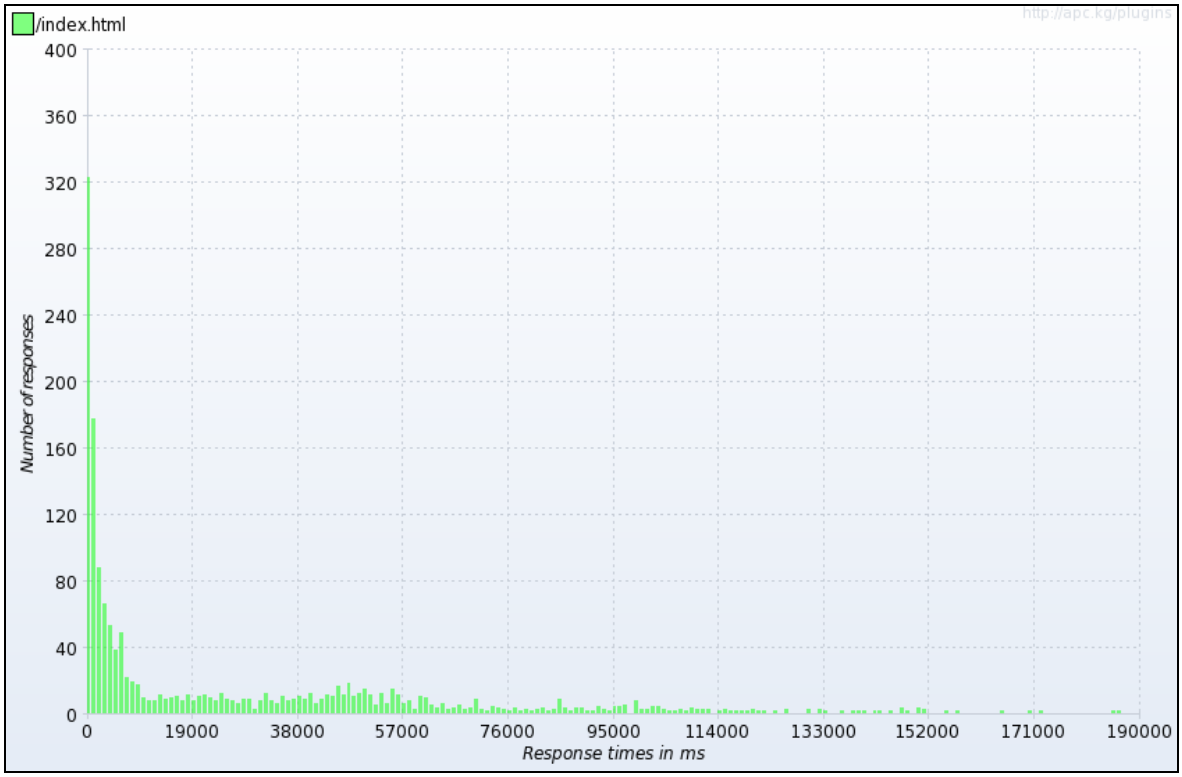
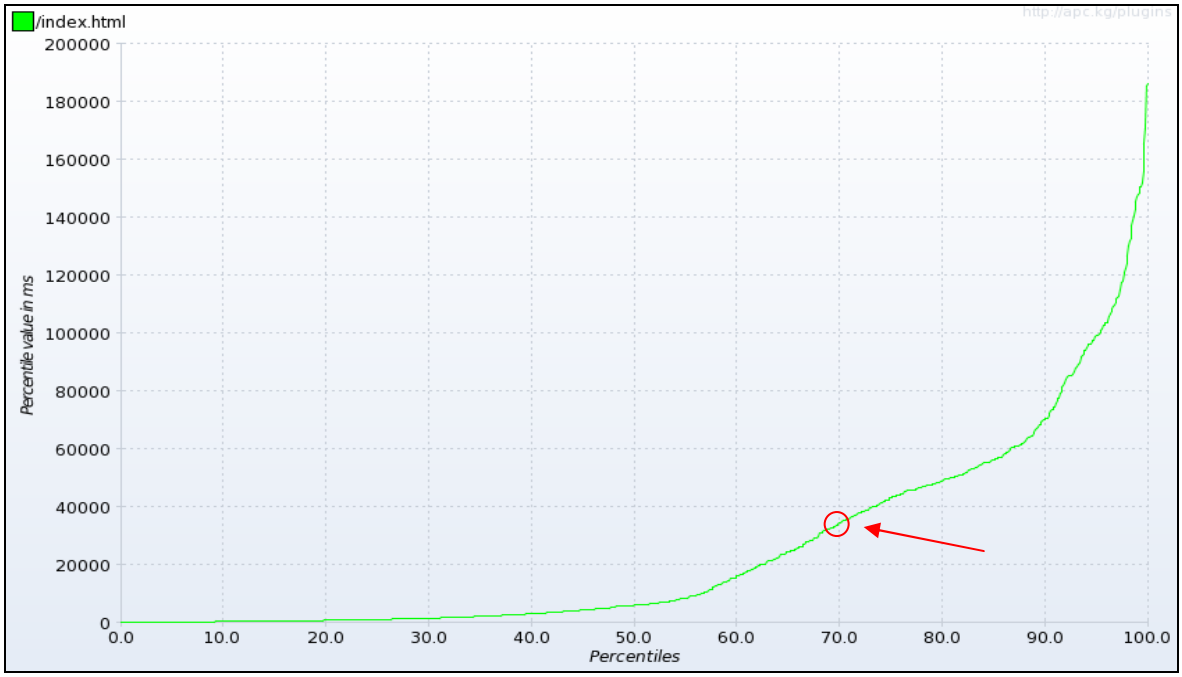


Figura 43 - Tempi di risposta in percentili per 300 richieste contemporanee



Dal risultato dei test descritti è emerso che il navigatore rispondeva alle esigenze ipotizzate dagli esperti del censimento che avevano peraltro fornito le seguenti stime sugli accessi previsti nell'ora di picco per il navigatore Ateco:

1. 26.000.000 famiglie;
2. 30% delle famiglie compilano sul *Web* ≅ 8.000.000;
3. 70% del punto 2 compilano il questionario negli ultimi 10 giorni utili ≅ 5.600.000;
4. 60% del punto 3 si avvalgono della funzione ricerca Ateco ≅ 3.400.000;
5. 12 ore per 20 giorni = 120 ore;
6. 3.400.000/120 ≅ 30.000 *query* l'ora;
7. 30.000/60 ≅ 500 richieste al minuto.

Dal grafico di Figura 41 si evince che con 100 utenti contemporanei, il 70% delle richieste ottiene una risposta entro i 10s. Dalla Figura 43 si vede invece che con 300 utenti contemporanei il 70% delle richieste ottiene una risposta entro i 40s. I tempi di risposta rimangono quindi accettabili con un numero di utenze che non superi i 100 utenti contemporanei. Considerando quindi un tempo di risposta di 10s per ogni richiesta, si può stimare un numero di 600 richieste risolte in 1 minuto. Questo dato era quindi al di sopra delle 500 richieste stimate conformemente alle previsioni sugli accessi.

9. Conclusioni

Le attività finalizzate alla gestione delle variabili testuali del 15° Censimento della popolazione hanno richiesto un notevole impiego di risorse specializzate sia negli aspetti tecnici che contenutistici delle Classificazioni di riferimento. Tuttavia si ritiene che tale sforzo abbia costituito un investimento sia nell'ottica di un elevato livello di qualità dei risultati ottenibili nel censimento stesso che in prospettiva per ulteriori applicazioni di interesse per l'Istituto.

Infatti, da un lato si è riusciti ad ottimizzare il processo di attribuzione dei codici alle risposte fornite a testo libero, tramite l'ampliamento delle basi informative messe a punto per ciascuna classificazione, nonché tramite la definizione di appositi iter procedurali per il trattamento delle risposte testuali al fine dell'individuazione dei codici corrispondenti, dall'altro si è riusciti a garantire un'uniformità di trattamento rispetto sia alla tecnica di acquisizione (questionario *Web* o cartaceo) che al soggetto preposto alla codifica.

Inoltre le basi informative aggiornate per il censimento potranno indubbiamente costituire un sostanzioso punto di riferimento non soltanto per la codifica di dati rilevati in altre indagini dell'Istituto, ma anche per gli utenti esterni che le potranno consultare sul sito istituzionale tramite i navigatori appositamente predisposti.

Riferimenti bibliografici

Cuccia F., De Angelis S., Laureti A., Macchia S., Mastroluca S., Perrone D. 2005. *La codifica delle variabili testuali nel Censimento generale della popolazione*, Documenti Istat (n.1/2005)

Ferrillo A., Istat S., Mazza L., Valery A., Vicari P. 2012. *La funzione su Web per l'individuazione del codice ATECO sulla base di una descrizione sintetica e monitoraggio delle performance*, Istat Working Papers, n.4 2012

Istat, 1958. *Anagrafe della popolazione*, Metodi e norme. Serie B – n. 3 edizione 1958

Istat, 1992. *Anagrafe della popolazione*. Metodi e norme. Serie B – n. 29 edizione 1992

Istat, 2007. *Census2010: Il progetto di aggiornamento delle Basi Territoriali di Census 2000.* Documento interno del Gruppo di lavoro finalizzato all'aggiornamento delle Basi Territoriali di Census 2000.

Istat, 2009. *L'ambiente di codifica automatica dell'Ateco2007*. Metodi e norme, 41. Roma

Istat, 2001. *Classificazione delle professioni,* Metodi e norme, Nuova serie - N.12, p.16, Roma

Jaro, M. A. 1989. *Advances in record linkage methodology as applied to the 1985 census of Tampa Florida*. Journal of the American Statistical Society. 84 (406): 414-420.

Macchia S., Mastroluca S., Reale A. 2001. *Planning the quality of the automatic coding process for the next Italian General Population Census.* Atti della Q2001 International Conference on Quality in Official Statistics (Stockholm, May 14-15, 2001)

Macchia S. e al. 2007. *Metodi e software per la codifica automatica e assistita dei dati*, Istat: Tecniche e strumenti n. 4. Roma

Mancini A. 2011. *"Il Censimento come investimento per il futuro delle statistiche demografiche e territoriali"*, 31° Convegno Nazionale ANUSCA, Riccione 17 novembre 2011

Winkler, W. E. 1990. *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*. Proceedings of the Section on Survey Research Methods (American Statistical Association). 354-359.

Zindato D. (a cura di) 2011. *15° Censimento generale della popolazione e delle abitazioni. Manuale della rilevazione*, Istat - ottobre 2011

Informazioni per gli autori

La collana è aperta ad autori dell'Istat e del Sistema statistico nazionale, e ad altri studiosi che abbiano partecipato ad attività promosse dal Sistan (convegni, seminari, gruppi di lavoro, ecc.). Da gennaio 2011 essa sostituirà Documenti Istat e Contributi Istat.

Coloro che desiderano pubblicare sulla nuova collana dovranno sottoporre il proprio contributo alla redazione degli Istat Working Papers inviandolo per posta elettronica all'indirizzo iwp@istat.it. Il saggio deve essere redatto seguendo gli standard editoriali previsti, corredato di un sommario in italiano e in inglese; deve, altresì, essere accompagnato da una dichiarazione di paternità dell'opera. Per la stesura del testo occorre seguire le indicazioni presenti nel foglio di stile, con le citazioni e i riferimenti bibliografici redatti secondo il protocollo internazionale 'Autore-Data' del *Chicago Manual of Style*.

Per gli autori Istat, la sottomissione dei lavori deve essere accompagnata da una mail del proprio dirigente di Servizio/Struttura, che ne assicura la presa visione. Per gli autori degli altri enti del Sistan la trasmissione avviene attraverso il responsabile dell'ufficio di statistica, che ne prende visione. Per tutti gli altri autori, esterni all'Istat e al Sistan, non è necessaria alcuna presa visione. Tutti i lavori saranno sottoposti al Comitato di redazione, che valuterà la significatività del lavoro per il progresso dell'attività statistica istituzionale. La pubblicazione sarà disponibile su formato digitale e sarà consultabile on line.

Gli articoli pubblicati impegnano esclusivamente gli autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat. Si autorizza la riproduzione a fini non commerciali e con citazione della fonte.